
THE UNIVERSITY *of* EDINBURGH

# Machine learning-based approaches for functional variant classification across mammals


**Rongrong Zhao**

A Thesis submitted for the degree of
*Doctor of Philosophy*
The University of Edinburgh
2023

# Declaration

I declare that the thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contributions and those of the other authors to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

The work presented in Chapter 3 was previously published in Communications Biology as *The conservation of human functional variants and their effects across livestock species* by Rongrong Zhao (student and author of the declaration), Andrea Talenti (supervisor), Lingzhao Fang, Shuli Liu, George Liu, Neil P. Chue Hong (supervisor), Albert Tenesa, Musa Hassan (supervisor) and James G. D. Prendergast (supervisor). This study was conceived by all of the authors. The contributions to the work were as follows: J.P. conceived the initial project idea, further developing it with R.Z., A.Ta. and M.H.. R.Z. and J.P. performed the majority of analyses with contributions from A.Ta., L.F., S.L., G.L., N.C.H. and A.Te.. L.F., S.L., G.L. and A.Te generated the cattle GTEx data used in this study. J.P. and R.Z. wrote the initial manuscript draft with all authors contributing to the final version.

Rongrong Zhao
September, 2023

# Acknowledgements

Firstly, I would like to express my gratitude to Professor James Prendergast, Dr Musa Hassan, Professor Neil Chue Hong, and Dr Andrea Talenti. I feel fortunate to have had all of you as my supervisors. Without your guidance, I wouldn't have been able to complete my PhD project and my thesis.

I want to extend my sincere thanks to my principal supervisor, Professor James Prendergast. Thank you for trusting me with this project and for your invaluable assistance throughout my PhD journey. Thanks for always being patience even when I came to your office with unexpected and sometimes trivial questions. I greatly appreciate all your amazing teaching and always being there offering me help. I would also like to thank my additional supervisor Dr Musa Hassan. I am deeply thankful for all your suggestions during our weekly meetings and for your help with my thesis. I would like to thank my other two additional supervisors, Professor Neil Chue Hong and Dr Andrea Talenti, for all your help during my PhD. I want to particularly thank Dr Andrea Talenti for your suggestions in addressing various bioinformatics challenges. I want to thank my thesis committee chair, Dr Vicky MacRae and Dr Alison Meynert, for all your helpful suggestions during my annual reviews.

I would like to thank all our lab members: Dr Rachel Owen, Shannon Massey, Siddharth Jayaraman, Dr Melissa Marr, Zexin Jiao, Ho Ching Brian Chan, Dr Jess Powell, Dr Lindsey Plenderleith and Dr Juliane Friedrich. Thank you for all the amazing presentations and discussions during our weekly group meetings, and I have learnt a great deal from all of you. I also want to express my gratitude to Professor Liam Morrison and Dr Tim Connelly for participating in our Monday meeting. Furthermore, I'd like to thank all the people at the Roslin Institute who have provided assistance and support during my PhD.

A special thank you to all the co-authors of my first paper: Dr Andrea Talenti, Dr Lingzhao Fang, Dr Shuli Liu, Dr George Liu, Professor Neil Chue Hong, Professor Albert Tenesa,

Dr Musa Hassan and Professor James Prendergast. Your contributions, especially those of Professor James Prendergast, were pivotal in the publication of my first paper. Additionally, I'd like to thank Dr Valentina Riggio for involving me in your research and collaborative paper.

Thank you to all my friends, especially Jingyi Ren, Chengyue Zhao, Heng Jiang, Ge Li, Li Zhang, for all your support during my PhD. Thank you all for being my "Internet Golden Retriever", always sharing funny stuff and listening to my concerns, even though we are scattered around the world and separated by different time zones. I would also like to thank all the friends I met in Edinburgh; you made my life in Edinburgh more enjoyable, especially during the pandemic.

Last but not least, a huge thank you to my mother and father for your unconditional support and encouragement. Thanks for always believing in me, being proud of me, and supporting my decisions.

# Abstract

As a result of the continued growth of the world's population, the demand for livestock products continues to grow. However, increasing livestock production results in more greenhouse gas emissions, and pressures on scarce resources such as potable water and land. Therefore, it is of vital importance to improve the productivity of livestock, including through advanced genomics breeding approaches and genome editing so that more can be produced without increasing animal numbers. The pivotal challenge of using advanced genomics breeding approaches is to identify the causal functional variants associated with the productivity traits of interest in livestock species. As in humans, genome-wide association studies (GWAS) have identified numerous genomic regions associated with diseases and traits in livestock, but it is difficult to determine the causal variants in these regions due to a range of factors such as linkage disequilibrium (LD). The overarching aim of my PhD was to utilize data-driven computational methods, such as machine learning, to improve the initial detection of novel functional variants in livestock species to ultimately enable the improvement of livestock breeding. This research focused on developing a reusable variant annotation pipeline for mammalian species with a broad range of features and demonstrating the utility of these features and machine learning approaches in predicting mammalian functional regulatory variants in both human and cattle.

Datasets suitable for machine learning are largely lacking in livestock. To address this and facilitate a diverse range of downstream projects I first developed a reusable variant annotation pipeline in Nextflow for use across platforms and species. The pipeline provides a wide range of annotations including sequence conservation, gene annotations, sequence context, and predicted functional genomic data from other machine learning tools such as Enformer, that can then be used in downstream variant analyses and employed as features in machine learning approaches for variant classification across species.

I first applied this pipeline to develop machine learning models for predicting where functional human variants have direct orthologues in livestock species, that may therefore be relevant to understanding livestock phenotypes. I demonstrate that it is possible to assign probabilities to whether a human variant will be found in other species from its annotations. Hundreds of human regulatory variants were identified with conserved functional impacts on gene expression in livestock species. This observation suggests it is possible to leverage information from well-annotated species, such as humans, to help with the prediction of regulatory variants and other functional variants in less well-annotated livestock species.

To explore the efficacy of using the annotation pipeline with machine learning approaches to predict functional variants, I applied them to directly predicting regulatory variants across humans and cattle. I compared the performance of various approaches of predicting cattle regulatory variants, including with or without incorporating annotations from humans. I highlight that the models incorporating human annotations and those based on cattle annotations demonstrated comparable performance, with the model relying on cattle annotations exhibiting a slightly superior performance.

Overall, the variant annotation pipeline and the machine learning models proposed in this thesis can be utilized to uncover the underlying characteristics of functional variants and prioritise functional variants related to important traits in livestock species for downstream genome editing or marker assisted breeding.

# Lay summary

Due to the on-going growth of the global population, there is a continuous increase in the demand for livestock products. However, this expansion of livestock production has brought about adverse environmental consequences due to various factors, including more greenhouse gas emissions, overgrazing, and increased water usage and pollution. Therefore, it is necessary to enhance the productivity of livestock, enabling us to obtain more livestock products without increasing animal numbers. This can be achieved through advanced genomics breeding approaches, which are based on the understanding of the livestock genome, especially the functional variants in the DNA sequence. DNA sequence is the fundamental unit found in living organisms, which is made up of four bases (A, C, G, T), carrying genetic information that determines the functioning of the organism. Within the same species, about 99.9% of the DNA sequence is the same across individuals, and the remaining variations can potentially result in the unique characteristics of an individual, such as height or predisposition to a disease. Among these variations, those that have a functional impact on biological processes and consequently affect downstream traits are called functional variants. Precisely identifying those functional variants associated with animal welfare and economically important traits, such as muscle mass and milk production, in livestock species is a crucial prerequisite for implementing advanced genomics breeding approaches.

Genome-wide association studies (GWAS) have identified genomic regions associated with specific traits or diseases. They have assigned a score to each variant in these regions, measuring the strength of the association between the variant and the target trait. However, it is difficult to identify the real causal functional variants from the large number of variants in these regions, mainly due to factors such as linkage disequilibrium (LD). In genetics, LD is a phenomenon where certain genetic variants that are close together on a chromosome tend to be inherited together more often than expected by chance. Some variants with significant scores may appear to be responsible for a trait, but in reality, they are only linked to the real causal variants because of LD. This poses challenges in accurately pinpointing the actual causal variant behind the traits of interest. Therefore,

some computational approaches, including machine learning, can be utilized to assist in the identification of functional variants. Machine learning algorithms are designed to find patterns from large amounts of data and distinguish different classes based on these patterns, making them suitable for solving problems like distinguishing functional variants from other variants.

In this thesis, I first developed a variant annotation pipeline to address the issue of datasets suitable for machine learning being largely lacking in livestock species. The pipeline provides a wide range of annotations, including distance from variants to important genomic elements and sequence context, which can reflect the characteristics of different variants. Next, the annotations from the pipeline were applied to develop machine learning models for predicting human functional regulatory variants that are shared across species, which may be relevant to understanding livestock phenotypes. Regulatory variants are functional variants that don't directly change proteins but can control how genes are turned on and off. Furthermore, the observation of hundreds of human regulatory variants found across species suggests that it is possible to leverage information from humans to aid in the prediction of regulatory variants in less well-annotated livestock species. Finally, I applied the annotation pipeline with machine learning approaches to directly predict functional regulatory variants across humans and cattle. I compared different approaches to predict cattle regulatory variants, including with or without incorporating human annotations, and highlighted their comparable performance in predicting cattle regulatory variants.

In summary, this work demonstrated the feasibility of utilizing the variant annotation pipeline and machine learning approaches to predict functional variants across species. The proposed tools can be employed to prioritise the functional variants associated with traits of interest in livestock species and further guide downstream genome editing and marker assisted breeding work.

# List of key abbreviations

| | |
|---|---|
| 3’ UTR | 3’ untranslated regions |
| 4d sites | Four-fold degenerate sites |
| 5’ UTR | 5’ untranslated regions |
| ANN | Artificial Neural Network |
| ARS-UCD1.2 | Latest *Bos taurus* reference genome |
| AS | Alternative splicing |
| ATAC-seq | Assay for Transposase-Accessible Chromatin Sequencing |
| AUROC/AUC | Area Under the Receiver Operating Characteristics curve |
| BCF | The Binary Variant Call Format |
| Btau_5.0.1 | An old version of *Bos taurus* genome assembly in 2015 |
| CAGE | Cap Analysis of Gene Expression |
| CatBoost | Categorical Boosting |
| CDS | Coding sequences |
| ChIP-seq | Chromatin Immunoprecipitation Sequencing |
| CNN | Convolutional Neural Network |
| CNV | Copy number variation |
| CPU | Central Processing Unit |
| CTCF | CCCTC-binding factor |
| DHSs | DNase I hypersensitive sites |
| DNase-seq | DNase I Hypersensitive Site Sequencing |
| DP | Sequencing depth |
| eQTL | Expression quantitative trait loci |
| FDR | False discovery rate |
| FN | False negative |
| FP | False positive |
| GBDT | Gradient Boosting Decision Trees |
| gEBVs | Genomic Estimated Breeding Values |
| GFF | General feature format |
| GPU | Graphics Processing Unit |

| | |
|---|---|
| GQ | Genotype quality |
| GTEx | Genotype-Tissue Expression project |
| GWAS | Genome-wide association studies |
| HAL | Hierarchical alignment format |
| hg38 | Latest human reference genome |
| KNN | K-nearest Neighbours |
| LD | Linkage disequilibrium |
| LDA | Linear discriminant analysis |
| LightGBM | Light gradient-boosting machine |
| lncRNA | Long non-coding RNA |
| MAF | Multiple alignment format |
| MCC | Matthew correlation coefficient |
| molQTL | Molecular quantitative trait loci |
| MPRA | Massively Parallel Reporter Assay |
| mRNA | Messenger RNA |
| ncRNA | Non-coding RNA |
| PCA | Principal component analysis |
| PIP | Posterior inclusion probabilities |
| PRO-cap | Precision Run-On Capturing |
| raQTL | Reporter assay quantitative trait loci |
| RNN | Recurrent Neural Network |
| ROC | Receiver Operating Characteristics |
| SHAP | SHapley Additive exPlanations |
| SNPs | Single Nucleotide Polymorphisms |
| SNVs | Single Nucleotide Variants |
| sQTL | Splicing quantitative trait loci |
| ss | Sufficient statistics format |
| SuRE | Survey of Regulatory Elements |
| SVD | Singular value decomposition |
| SVM | Support Vector Machine |
| SVs | Structural variants |

| | |
|---|---|
| TF | Transcription factor |
| TN | True negative |
| TP | True positive |
| TSS | Transcription start site |
| VCF | The Variant Call Format |
| VEP | Variant Effect Predictor |
| WGS | Whole-genome sequencing |
| XGBoost | Extreme Gradient Boosting |

# List of Figures

# List of Tables

# Table of Contents

# Chapter 1 Introduction

With the on-going growth of the world's population and the improvement of worldwide living standards, the demand for livestock products is increasing rapidly, from food products such as dairy and meat, to non-food products such as wool and leather. Furthermore, livestock contributes not just to food and nutrition but also to some social functions in many developing countries, such as promoting gender equality by enabling women to possess livestock. (Swanepoel et al., 2010). However, increasing livestock production leads to more greenhouse gas (GHG) emissions, which is the main cause of global climate change. In turn, effects of climate change, such as hotter temperatures, increased drought, etc., have been proved to have negative impacts on livestock performance (Cheng et al., 2022; Rojas-Downing et al., 2017). Therefore, increasing livestock production per animal, reducing the need for increased animal numbers, is of critical importance to a sustainable livestock production system.

Conventional selective breeding methods have improved livestock productivity, but are constrained by the standing genetic variation in specific species as well as the spontaneous de novo mutations within each generation (Tait-Burkard et al., 2018). Advanced genomic approaches, such as genome editing, could be used to overcome the limitations of such conventional breeding methods and improve livestock genetics more efficiently. For example, studies have found that an underactive myostatin gene (*MSTN*) can lead to increased muscle growth in cattle, sheep and goat, which is an important livestock trait of economic significance (Tait-Burkard et al., 2018). The underlying genetic variations sit directly within the target gene and introducing the desired mutations in *MSTN* into cattle, sheep and goat can eventually result in significantly more muscle mass (Tait-Burkard et al., 2018). However, *MSTN* mutations resulting in double muscling may have implications beyond benign outcomes, such as dystocia in cattle (Yang, 2014). These challenges could potentially be circumvented with a deeper understanding of refining genome editing approaches. In this scenario, regulatory variants could offer greater utility.

A key challenge of using advanced genomics breeding approaches is to identify the causal functional variants associated with the downstream phenotypes. Some studies, such as genome-wide association studies (GWAS), have identified loci associated with specific diseases and traits in both human and livestock species (Schaid et al., 2018). However, the identification of the causal variant is still challenging as there are most often plenty of variants in the identified loci above the threshold for genome-wide statistical significance (commonly p-value < $5 \times 10^{-8}$ in human studies) due to linkage disequilibrium (LD) (Schaid et al., 2018). Fine-mapping is the method used for elucidating the variants that are most likely to be functional for complex traits (Broekema et al., 2020). However, the effect sizes of the variants on the traits, the LD structure and some other factors can impact the performance of fine-mapping approaches (Schaid et al., 2018). Moreover, most of the current work on identifying causal variants using fine-mapping approaches has focused on humans, for example through the Genotype-Tissue Expression project (GTEx). Therefore, utilizing state-of-the-art computational methods such as machine learning approaches for improving the initial detection of novel functional variants in livestock species is necessary.

This chapter will be divided into two parts. First, I will introduce different types of variants and then delve into the consequences of functional variants. I will also discuss the current approaches for prioritizing and identifying functional variants. Second, I will introduce different types of basic machine learning algorithms, explain the general process of machine learning approaches, and then review the applications of machine learning in genomics.

## 1.1 Genomic variants

According to the Human Genome Project, humans are 99.9% identical at the base-pair level (nucleotides: A, G, C, T) across individuals (Collins & Mansoura, 2001). Genomic variants, differences in DNA sequence among individuals, are a major contributor to our unique characteristics such as hair colour or predisposition to a disease. These variants can have varying effects, with some causing significant changes and others influencing traits in combination with other genetic and environmental factors (Health (US) & Study,

2007). In livestock species, genomic variants can not only influence agriculturally important traits like milk production but also affect welfare relevant traits such as diseases resistance. The impact of variants, harmless, helpful or hurtful, depends on when, where and how the variant modifies the original genetic makeup. Understanding the function of genomic variants is vital for unraveling the genetic basis of complex traits. While Mendelian inheritance offers a fundamental framework for genetic diseases resulting from single genes with distinct effects, inheritance patterns for many traits are more intricate. To better understand the mechanisms under different phenotypes or diseases in mammals, it is necessary to know different types of variants and more importantly, their consequences.

### 1.1.1 Variant classification

According to the Ensembl genome database (https://www.ensembl.org), variants can be classified mainly into two categories, sequence variants and structural variants according to their impact range of the sequence (Cunningham et al., 2022).

#### 1.1.1.1 Sequence variants

Sequence variants can be further divided into five categories as shown in Table 1.1.

**Table 1. 1 Different types of sequence variants and the corresponding examples**

| Variant Type | Description | Example |
|---|---|---|
| **SNP** | Single Nucleotide Polymorphism | **Ref**: ...TCGATAT... **Alt**: ...TCGGTAT... |
| **Insertion** | Insertion of one or more nucleotides | **Ref**: ...TCGATAT... **Alt**: ...TCGATGTAT... |
| **Deletion** | The opposite of insertion, deletion of one or more nucleotides | **Ref**: ...TCGATGTAT... **Alt**: ...TCGATAT... |
| **Indel** | Insertion-deletion, an insertion and a deletion, affecting two or more nucleotides | **Ref**: ...TCGATAT... **Alt**: ...TCGAGCTTAT... |
| **Substitution** | A sequence alteration in which the length of the variant change matches that of the reference | **Ref**: ...TCGATTAT... **Alt**: ...TCGGGTAT... |

- **Single Nucleotide Polymorphism (SNP):** SNPs are the most common genetic variants in the genome (Nelson et al., 2004). They are the variations that occur at

single base position within genes or noncoding regions in the genome with appreciable frequency in the population (> 1%) (Brody, 2016). As shown in Table 1.1, the example SNP is a replacement of the nucleotide A in the reference genome with the nucleotide G as an alternative allele. It should be noted that, single nucleotide variant and single nucleotide polymorphism are not interchangeable. The variant must be with the frequency of at least 1% in the population to be qualified as a SNP. Due to natural selection, SNPs more frequently appear in non-coding regions than in coding regions (Barreiro et al., 2008). SNPs are widely used as markers to specific regions in the genome which may be associated with traits of interest, such as is often the case in genome-wide association studies (Uffelmann et al., 2021).

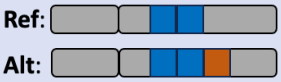
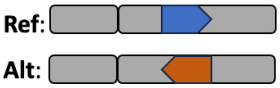
- **Insertion:** Insertion refers to variants where one or more extra nucleotides are inserted into the genome relative to the reference genome. which usually happen during DNA replication (T. A. Brown, 2002).
- **Deletion:** Deletions, the opposite of insertions, are where consecutive bases in the reference genome are missing from an individual. Insertions and deletions are highly abundant in our genome, only second to SNPs in terms of their numbers (Mullaney et al., 2010).
- **Indel:** Although often used as a catch all term to cover both insertions and deletions, insertion-deletion (indel) also covers where deletion and insertion mutations co-occur in the genome (Sehn, 2015). As shown in the example part of Table 1.1, base A is deleted and then three bases GCT are inserted into the corresponding position.
- **Substitution:** Substitutions are a type of mutation where one or more nucleotides are replaced by other nucleotides of the same length (Sehn, 2015).

#### 1.1.1.2 Structural variants

Structural variants (SV) refer to large genomic alterations generally more than 50 base pairs in size (Mahmoud et al., 2019). Compared to general sequence variants, structural variants are more difficult to detect because of their structural complexity (Sudmant et al., 2015). As shown in Table 1.2, structural variants can be divided into three subtypes: Copy number variation (CNV), inversion and translocation.

- **CNV:** A large part of our genome is made up of repeating sequences and the number of copies of some specific segments varies among individuals. This kind of variation is called copy number variation (CNV), which is usually caused by unequal recombination (Sudmant et al., 2015).
- **Inversion:** Inversions are structural variants in which the orientation of some DNA fragments are reversed relative to their ancestral orientation on the chromosome (Escaramís et al., 2015).
- **Translocation:** Translocations are where a region of nucleotide sequence changes its position in the genome without loss or gain of the genetic materials. This can happen both intrachromosomally or interchromosomally (Escaramís et al., 2015).

**Table 1. 2 Different types of structural variants.**

| Variant Type | Description | Example |
|---|---|---|
| **CNV** | Copy Number Variation |   |
| **Inversion** | A continuous nucleotide sequence is reversed at the same position |   |
| **Translocation** | A segment of a nucleotide sequence that has shifted to a different location |   |

## 1.1.2 Consequences of variants

Although large amount of variants have been found in human and other species, the majority of them are tolerated and have no impact on downstream phenotypes, such as fourfold degenerate variants that change the DNA sequence of genes but do not affect the protein's product because of the redundancy of the genetic code. These variants are non-functional variants while those that affect the molecular function of a protein are functional variants, including variants linked to diseases and traits (Gonzalez-Perez et al., 2013). Therefore, studying functional consequences of variants can help us understand the underlying molecular mechanisms that link genotype to phenotype. Sequence variants can result in functional consequences by altering the protein structure, changing

post-translational modification sites, etc., while the functional consequences of structural variants can be difficult to describe by biochemical effects and they are typically defined by general phenotypic traits. Here, I will focus on introducing the consequences of sequence variants.

Figure 1.1 (A) shows calculated variant consequences from the Ensembl database in different genomic elements. Figure 1.1 (B) categorizes these consequences according to their impact severity assessed by the Ensembl impact rating system, which relies on SnpEff, a tool for variant effect prediction (Cingolani et al., 2012; Cunningham et al., 2022). High impact means variants are assumed to have a highly disruptive impact, such as causing the loss of protein function via protein truncation. Moderate impact indicates that the variants are less disruptive but may change the effectiveness of proteins. Low impact consequences refer to the variants that are most likely to be harmless. There is another modifier category which includes variant consequences with impacts that are difficult to predict or for which there is no evidence of impact severity. This section focuses on variant consequences predicted by Ensembl, specifically the effects of the variants according to their locations in the genome (Sections 1.1.2.1 and 1.1.2.2). However, it is important to note that variant effects can be more intricate. In Section 1.1.2.3, I will explore these complexities, including polygenic inheritance and penetrance.

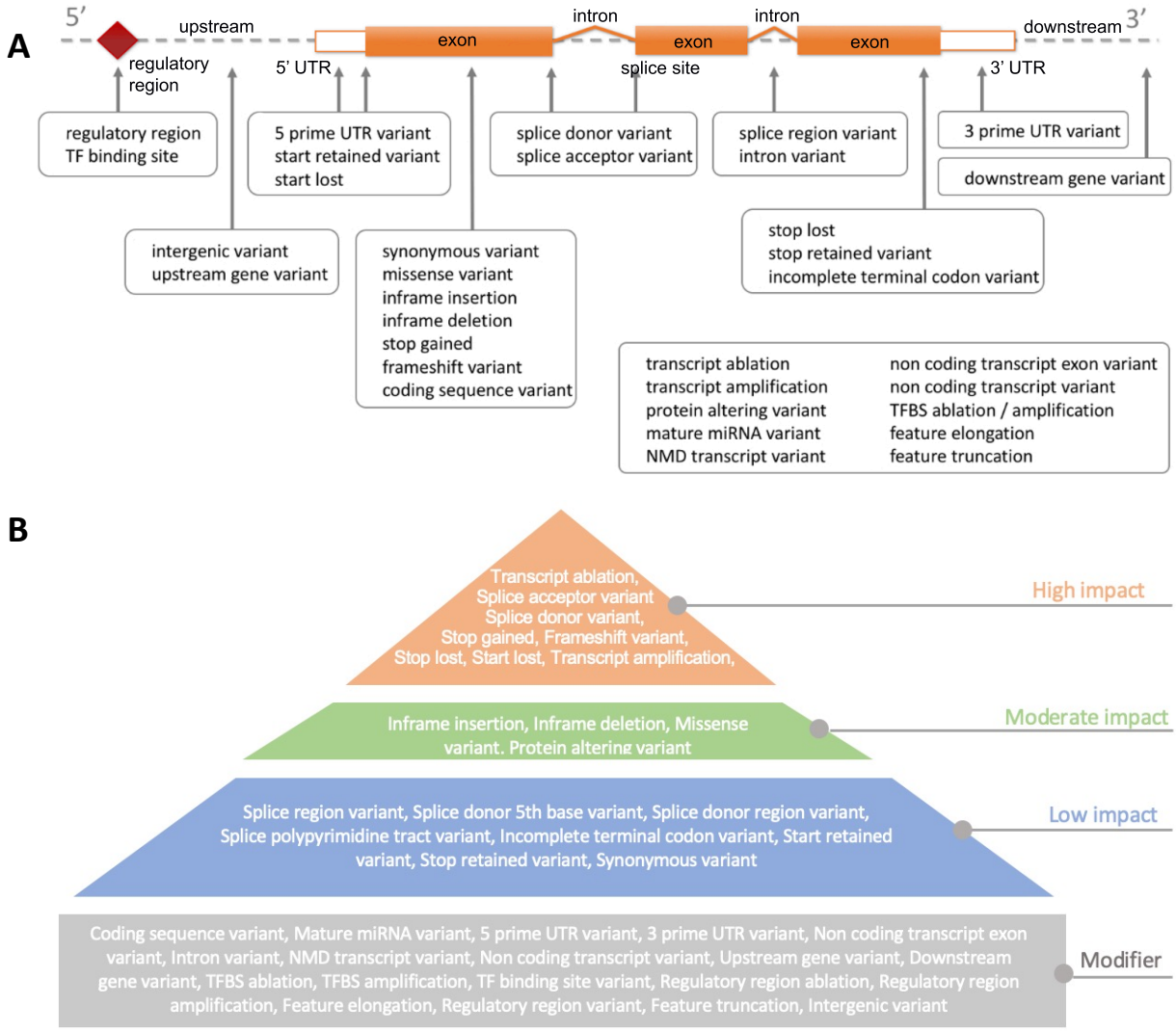



**Figure 1.1 Calculated variant consequences.** (A) Gene model and possible variant consequences in different regions. The variants are described using terms adapted from sequence ontology (SO) terms (Eilbeck et al., 2005a). Figure adapted from the Ensembl website. (B) The estimated severity of the consequences according to Ensembl impact rating system, from high impact to low impact (Cunningham et al., 2022). The modifier category includes the variant consequences where the impact is difficult to predict or there is no evidence of impact severity.

#### 1.1.2.1 Variant effect in coding regions

Gene expression includes two processes: transcription and translation. The information in DNA is transferred to mRNA through transcription and the mature mRNA is then translated into protein. In the translation process, mRNA is "translated" according to the three-base amino acid code, which links DNA sequence to the corresponding protein amino acid sequence. Therefore, the effect of variants that fall within the coding regions depends on their effect on the genetic code.

- **Synonymous variant:** Synonymous variants, also called silent variants, are the mutations that do not change the produced amino acid sequence due to the redundancy of the genetic code (T. A. Brown, 2002). Generally, synonymous variants have low or no impact on the synthesized proteins. However, evidence has shown that some of these silent mutations can impact mRNA stability or translation efficiency, ultimately altering protein functions in mammals (Chamary & Hurst, 2005; Kimchi-Sarfaty et al., 2007). For example, a synonymous variant in the Multidrug Resistance 1 (MDR1) gene has been found to influence the downstream multidrug resistance protein, resulting in altered drug and inhibitor interactions (Kimchi-Sarfaty et al., 2007).
- **Nonsense variant:** A nonsense variant will lead to a premature stop codon that results in a partially or completely nonfunctional protein (J. Sharma et al., 2020). This type of variant includes stop gained, stop loss, start loss, etc. The majority of nonsense variants have severe impact as shown in Figure 1.1 (B). For example, a nonsense variant in a gene which encodes dystrophin protein can cause a genetic disease called Duchenne muscular dystrophy (Flanigan et al., 2011). Another example is a nonsense variant found in the sheep coiled-coil domain containing 65 (*CCDC65*) gene that can induce a premature stop codon which eventually causes respiratory failure in sheep (Ben Braiek et al., 2022).
- **Missense variant:** Missense variants alter the original genetic code which then results in a different amino acid (Zhang et al., 2012). Some of the missense variants lead to altered amino acids that have similar properties to the original amino acids. These are conservative missense variants, as the function of the produced proteins may remain similar. Non-conservative missense variants result in amino acids with different properties and the function of the protein may change. For example, a

missense mutation in the heat shock protein family B 7 (*HSPB7*) gene is presumed to be associated with heat tolerance in indicine cattle (L. Zeng et al., 2019).

- **In-frame deletion, in-frame insertion and frameshift variants**: These kinds of variants are indels in coding regions. An indel whose length is divisible by three, resulting in amino acid insertion or deletion within the protein, is referred to as an in-frame insertion or deletion (Douville et al., 2016). The impact severity of in-frame variants depends on the location and size of the variants. An example of this is the location difference of the in-frame variants in the Duchenne muscular dystrophy (DMD) gene that can result in different disease phenotypes, some are mild while others are severe (E. M. Gibbs et al., 2019). If the indel length is not a multiple of three, it is a frameshift variant which will change the gene's reading frame (Douville et al., 2016). The improper translation result will ultimately lead to a defective or abnormal protein which may be completely new or nonfunctional. Therefore, frameshift variants typically have a high impact, especially when the indel occurs early in the sequence.

#### 1.1.2.2 Variant effects in non-coding regions

As discussed in the previous section, the functional role of variants in coding regions is relatively clear as they directly affect the genetic code and the protein product. Non-coding regions refer to the genomic regions that do not directly code for amino acids, which account for about 98-99% of the DNA in mammals (Shabalina & Spiridonov, 2004). The biological function of non-coding regions is still not fully understood but some of them are found to be responsible for controlling gene expression. Although variants falling within noncoding regions do not alter proteins directly, they could potentially affect downstream phenotypes by altering these regulatory elements such as enhancers, silencers etc. (Schipper & Posthuma, 2022). As shown in Figure 1.1 (B), the impact of noncoding variants is usually difficult to predict due to their indirect effect and the lack of knowledge of the corresponding non-coding regions. In this part, I will discuss the possible impact of variants in non-coding regions.

- **Regulatory element variants**: Regulatory elements include those specific regulatory sequences that generally bind with transcription factors (TFs) and that impact the

transcription process, such as promoters, enhancers, silencers and insulators. Figure 1.2 shows how these elements work to regulate gene expression.

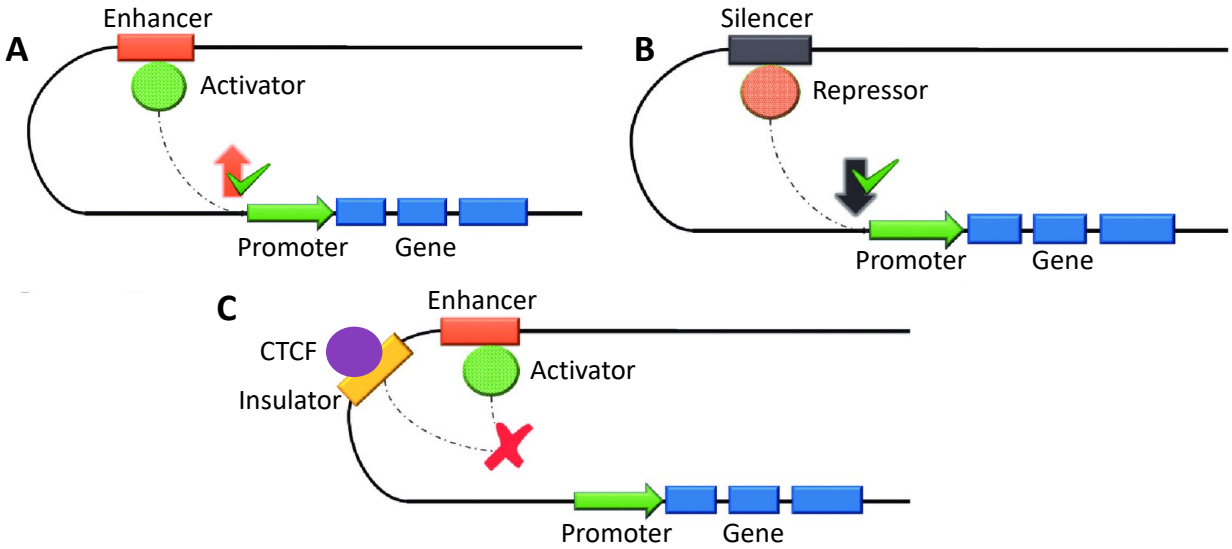



**Figure 1.2 Effects of regulatory elements.** (A) Enhancers activate gene promoters by binding to a protein called an activator. (B) Silencers bind to repressor proteins to prevent gene expression. (C) Insulators can bind to a transcriptional repressor called CCCTC-binding factor (CTCF) to prevent ectopic enhancer-promoter interaction. Figure adapted from (Rojano et al., 2019).

Enhancers are regions that bind to a DNA-binding protein called an activator and regulate gene expression by dictating the cell type, expression timing and expression level of the gene. They can sit close to the regulated gene and can also be located a megabase or more from the regulated gene or even on another chromosome. A gene could potentially be regulated by multiple enhancers and one enhancer is able to regulate different genes (Karnuta & Scacheri, 2018). Genetic variants that disrupt enhancer elements are therefore able to change when, where, and how genes are expressed. A representative example of an enhancer variant is one found in the enhancer that regulates sonic hedgehog (*SHH*). The misexpression of *SHH* protein caused by this variant results in preaxial polydactyly (PPD), a common limb malformation observed in humans, mice, cats and chickens (Anderson et al., 2012). Nevertheless, studies have found that multiple enhancers with a similar function can locate near to the same gene to promote phenotypic robustness, which means that a

single loss-of-function enhancer variant will not necessarily change gene expression (Osterwalder et al., 2018). This makes predicting the impact of such variants harder.

Variants in promoters and silencers have also been identified with an impact on phenotypes and associated with disease. For example, a gain-of-function regulatory variant was identified which results in the generation of a new promoter-like element that disrupts the normal activation of alpha globin-like genes and eventually cause alpha thalassemia (De Gobbi et al., 2006). Another example is the identification of a variant in a putative silencer region of dog retinal degeneration associated genes *EDN2* and *COL9A2* (Kaukonen et al., 2020). Silencers are sequences that bind to transcription regulation factors called repressors and prevent the transcription from DNA to RNA. This silencer mutation disrupts the repression function of the silencer and causes the overexpression of these two genes in the dog leading to progressive retinal atrophy (PRA) (Kaukonen et al., 2020).

- **Non-coding RNA (ncRNA) gene variants**: ncRNAs are functional RNA molecules that do not encode proteins directly but can impact protein-coding genes through regulating RNA production, translation or degradation (Rojano et al., 2019). Several mechanisms can be involved in this process such as splicing, protein binding, etc. The most well-known type of ncRNAs is microRNAs (miRNA), which are short RNA molecules containing less than 25 nucleotides. They are responsible for gene silencing. The preprocessed mature miRNA is loaded into an RNA-induced silencing complex (RISC) which can either repress translation or cleave protein-coding transcripts (mRNA) to regulate the downstream translation (Esteller, 2011). An example is that Bourdon et al. identified thousands of miRNA variants associated with dairy traits in bovine, caprine, and ovine species (Bourdon et al., 2021).

  Another typical class of ncRNA is long non-coding RNAs (lncRNA) which usually contain more than 200 nucleotides and that can be responsible for alternative splicing (AS) and some other biological processes (Esteller, 2011). Alternative splicing is an important cellular process for the generation of mature mRNAs from precursor RNAs.

Studies have shown that some variants in lncRNA could potentially affect alternative splicing processes and therefore lead to the occurrence and development of cancer (Ouyang et al., 2022).

- **5' UTR and 3' UTR variants**: 5' and 3' untranslated regions (UTR) are non-coding mRNA regions that take charge of crucial post-transcriptional regulation processes such as governing RNA stability and translation efficacy (Schuster & Hsieh, 2019). 5' UTRs usually contains a Kozak consensus sequence (ACCAUGG), which includes the initiation codon for translation, and other regulatory sequences such as CpG sites (Steri et al., 2018). Besides, the secondary structures occurring within 5' UTRs are also important for translation regulation. Therefore, variants that fall within the 5' UTR of genes can impact RNA stability, translation and consequently affect the protein product. An example of this is the variant in the Kozak sequence of the 5' UTR of the β-globin gene, which is relevant to β-Thalassemia, resulting in about 30% reduction of the translational rate of this gene (Steri et al., 2018). 3' UTRs sit downstream of the coding sequence and are also responsible for RNA stability and mRNA translation. In addition, 3' UTRs can be characterized by miRNA' binding sites, therefore, variants in the 3' UTR may disrupt miRNA binding and further impact gene expression.

### 1.1.2.3 Complexity of variant consequences

Sections 1.1.1 and 1.1.2 introduced different types of variants and their predicted consequences based on their positions. While this offers a helpful framework, it is important to acknowledge that the effects of variants on phenotypes can be more intricate than a simple cause-and-effect relationship.

Mendelian inheritance serves as a foundational framework for understanding genetic diseases resulting from mutations in single genes with clear-cut effects. However, many traits, especially complex diseases and some common physical characteristics such as height or weight, are not solely determined by single genes (Lappalainen et al., 2024). The combined effect of multiple genes, each with a subtle individual contribution, complicates the identification of the causal variants.

Furthermore, the presence of a particular variant doesn't always guarantee the manifestation of the associated trait. Genetic penetrance describes the likelihood that an individual carrying a specific genetic variant (genotype) will develop the associated trait (phenotype) (Koellner et al., 2018). Incomplete penetrance occurs when some carriers of a variant don't exhibit the associated trait, potentially due to the influence of other genetic or certain factors such as age (Koellner et al., 2018). For example, 75% of individuals exhibit symptoms of Huntington disease at age 65, a much higher percentage than the 25% observed at the age of 50 (Parsons & Raymond, 2015).

Dominance describes how different alleles at the same locus on chromosomes interact to influence the resulting phenotypes (Billiard et al., 2021). Mendelian inheritance typically describes complete dominance, where the heterozygous individual shows the same phenotype as one homozygous parent (Miko, 2008). However, dominance patterns can be more complex, including incomplete dominance and codominance. In incomplete dominance, the heterozygous genotype exhibits a phenotype that is intermediate between the homozygous genotypes, while in codominance, both alleles are expressed in the phenotype (Billiard et al., 2021). An example of incomplete dominance is observed in horse coat colour, where different combinations of alleles at the cream locus result in varying horse coat colours (Thiruvenkadan et al., 2008).

In addition to these key factors, some other factors beyond the DNA sequence itself can also impact the consequences of variants. For example, the effects of a variant can be influenced by environmental exposures such as certain chemicals (Virolainen et al., 2023). Understanding these complexities is crucial for effectively utilizing genome-wide association studies (GWAS) to identify genetic variants associated with complex traits or diseases. The next section will delve into the principles and approaches used in GWAS.

### 1.1.3 From genome-wide association study (GWAS) to causal variant

The completion of the human genome project and the population-level survey of genetic variants via the international HapMap project provided the foundation of further studies on understanding how genomic variants contribute to common diseases. Utilizing the

reference genome from the human genome project and the map of genetic variants from HapMap, together with other new technologies in genotyping, genome-wide association studies (GWAS) were developed to test genome-wide sets of genetic variants and find out those associated with specific diseases or traits. The first GWAS, published in 2005, identified an association between a variant in the complement factor H gene (*CFH*) and age-related macular degeneration (AMD) (Klein et al., 2005). With the completion of several livestock genome assemblies, such as the cattle whole genome assembly published in 2009 (A. Sharma et al., 2015), and the emergence of economical genotyping assays such as Illumina arrays, it became feasible to expand GWAS to livestock species. For example, an initial study in livestock GWAS examined variants associated with several cattle growth traits, such as weight and height (A. Sharma et al., 2015).

To date, researchers have identified thousands of variants associated with diseases or other complex traits in both human and livestock species (Uffelmann et al., 2021). Although structural variants and other sequence variants can be considered in GWAS, their focus is on SNPs. The typical outputs of GWAS are genomic risk loci, which include correlated SNPs where the occurrence of their genotypes in individuals are significantly associated with traits of interest. These results can then be used in post-GWAS analysis for observation and understanding of causal variants and other biological mechanisms underlying the traits of interest. Before introducing the general steps in GWAS, this section will first discuss some important concepts underlying GWAS.

#### 1.1.3.1 Linkage disequilibrium and haplotype block

One of the important characteristics of SNPs is linkage disequilibrium (LD). Some alleles occur together more often than expected by chance and this usually happens when these alleles are physically close on the DNA (Slatkin, 2008). This kind of non-random co-inheritance of the alleles at different loci in the genome is called LD (Slatkin, 2008), which is caused by multiple genetic factors including selection, recombination rate, mutation rate. A cluster of linked SNPs with strong LD is called a haplotype block, encompassing alleles that are inherited together with little opportunity for contemporary recombination (Wall & Pritchard, 2003).

Haplotype blocks are useful in association studies. Compared to using individual SNPs in association studies, which leads to large numbers of individual variants to test, using haplotypes can be more robust. The genome is partitioned into chunks according to the haplotype blocks and these chunks are used as units in association studies which reduces the multiple hypothesis testing burden and can therefore increase the ability to detect significant loci. Furthermore, utilizing LD between SNPs in association studies can also be a complement to the limitations of the genotyping approaches. As shown in Figure 1.3, there is an actual causal variant directly associated with the phenotype of interest in the candidate gene that we are unable to genotype. Meanwhile, there exists a SNP marker in LD with the causal variant that can be genotyped. By detecting the indirect association between the genotyped SNP marker and the phenotype, this haplotype block that contains the causal SNP and causative gene can be mapped. Consequently, the underlying mechanisms of the phenotype can be analysed in the association study.

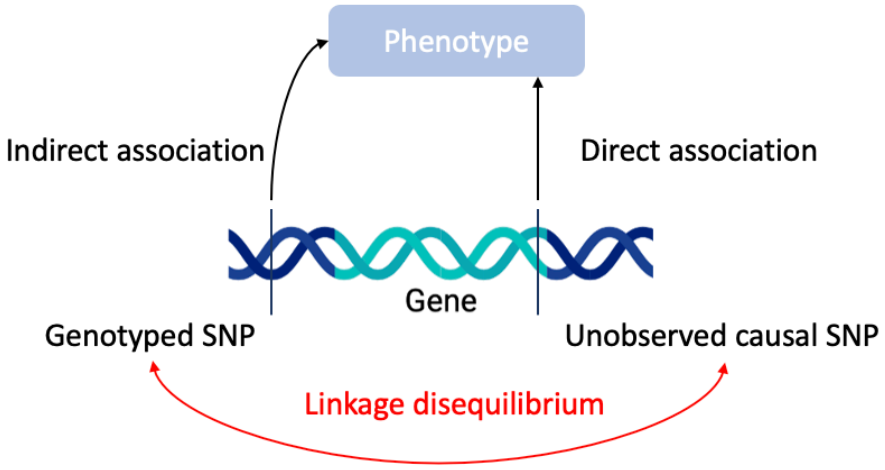



**Figure 1.3 The presence of LD results in both direct and indirect associations between the variants and the phenotype in the haplotype block.**

### 1.1.3.2 Overview of steps in GWAS

The first step of GWAS involves study design and data collection. The study can be set up as a case-control study if the traits of interest are binary (e.g. disease or no-disease), or as a study using quantitative measurements if the traits are quantitative (e.g. height). As shown in Figure 1.4 (A), individuals are then genotyped using microarray-based

approach or whole-genome sequencing (WGS). Microarray-based approaches are commonly used for GWAS due to their efficiency and low-cost. Compared to microarray-based approach, WGS can get almost every genotype of a complete DNA sequence, which makes it an ideal genotyping method. However, it is computationally intensive (Uffelmann et al., 2021). The called variants from genotyping, along with additional information such as anonymous individual ID, phenotype information, genotype batch information, are used as inputs for the GWAS analysis.

To generate reliable results, the next step after genotyping is the quality control work. This includes filtering missing SNPs, removing genotyping errors, correcting population stratification, etc. Of all the quality control work, population stratification correcting is crucial. Population stratification refers to the systematic ancestry differences between control and case groups, which can lead to spurious associations between traits and genetic variants in the absence of appropriate correction (Price et al., 2010). Several software can be used for quality control, for example, PLINK can be used for SNP filtering and many other key steps, SMARTPCA is useful for principal component analysis and further correction of population stratification. Having undergone quality control, variants usually go through phasing and imputation in the next steps. Phasing is the process of deducing whether the genotyped alleles are from the maternal or paternal haplotype and imputation is the statistical inference of the missing genotypes that are not assayed from genotyping step (Uffelmann et al., 2021). The HapMap (R. A. Gibbs et al., 2003) or 1000 Genomes Consortium (Auton et al., 2015) haplotypes are often used as the reference panel for genotype imputation. This imputation helps fill in gaps where a genotype is unknown, improving the identification of the causal variants and facilitating the post-GWAS analysis.

After quality control and imputation, the subsequent processes include association testing. GWAS typically utilize regression analysis, a statistical test that measures the association between traits and variants by assessing the relationship between a dependent outcome variable and independent variables. Depending on the types of traits, binary or quantitative, different models can be used for association testing. For the binary traits, the

association between variants and traits is commonly measured using logistic regression models. For quantitative traits, the association can be tested with multiple linear regression models. Logistic regression models are indeed the extension of linear regression models, in which logistic functions are used to transform a linear model output to probability of a case status occurring at a given genotype (Bush & Moore, 2012). Furthermore, covariates such as sex, age, and ancestry principal components, are commonly included in the regression models to control for such confounding effects when studying some common diseases (Uffelmann et al., 2021).

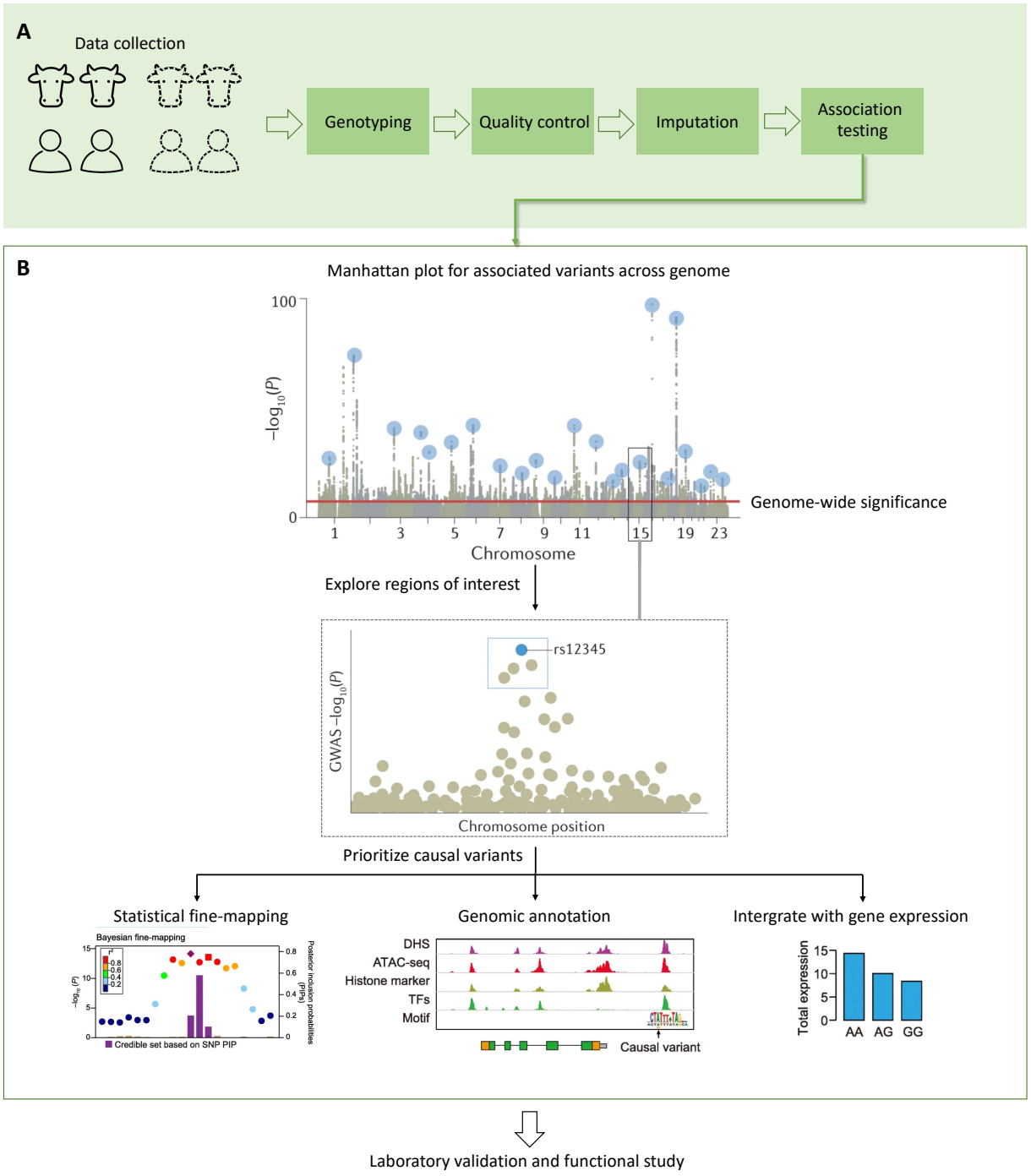



**Figure 1.4 Typical steps from GWAS to causal variant prioritization.** (A) A typical GWAS workflow: from data collection to statistical association testing. (B) Steps for prioritizing causal variants for the trait of interest after the initial GWAS work. The Manhattan plot shows the positions of associated variants and their corresponding association strengths. Statistical fine-mapping approaches are applied to identify the possible causal variants. Genomic annotations and

expression quantitative trait loci (eQTLs) mapping are used for exploring the underlying functional impact of the variants. Figure adapted from (Rao et al., 2021; Uffelmann et al., 2021).

### 1.1.3.3 Multiple hypothesis testing and p-values

In GWAS, researchers usually do statistical testing on each SNP marker and use p-values to measure the statistical significance of an association between a genetic variant and a trait. Statistical significance here means an assessment of the null hypothesis of no association between the variant and phenotype. A general framework of hypothesis testing in GWAS first defines a null hypothesis ($H_0$) and an alternative hypothesis ($H_1$). The null hypothesis states that the variant is independent of the phenotype whereas the alternative hypothesis is the variant is associated with the phenotype. Suppose there is a random variable $X$ with range $X$, $R \subset X$ is the rejection region. If $X \in R$, then the null hypothesis is rejected, otherwise the null hypothesis is accepted. The p-value is the probability of obtaining a test statistic at least as extreme as what has been actually observed with the assumption of the null hypothesis being true (Bush & Moore, 2012). In other words, a p-value in GWAS indicates the chance of seeing the result given the hypothesis that the variant is not associated with a trait. Therefore, a small p-value can be taken as evidence of the null hypothesis being incorrect. The possible outcomes of the hypothesis test are shown in the following table. A type I error occurs when $H_0$ is wrongly rejected whereas type II error is when $H_0$ is wrongly accepted.

**Table 1. 3 Decision table for hypothesis test**

| | | **Reality** | |
|---|---|---|---|
| | | **$H_0$ True** | **$H_1$ True** |
| **Decision** | **Reject $H_0$** | Type I error (False positive) | Correct |
| | **Accept $H_0$** | Correct | Type II error (False negative) |

Researchers use a predefined $\alpha$ value as the p-value threshold to determine the level of statistical significance required to have confidence in the authenticity of the association. When the p-value is smaller than $\alpha$, the null hypothesis is rejected, otherwise it is accepted. Traditionally, this $\alpha$ is predefined as 0.05. This indicates that the null hypothesis, where the variant is independent of the phenotype, is rejected 5% of the time when it is actually true. In other words, there is a one-in-twenty chance of the association being a false positive. However, this threshold is too lenient for GWAS since GWAS can involve millions of tests. A threshold of 0.05 may lead to over 100,000 false positives, which means over 100,000 spurious associations are considered as statistically significant. Such a high false positive rate is unacceptable.

To avoid this problem, some approaches use multiple testing correction in GWAS. A simple correction is the Bonferroni correction which adjusts the threshold to 0.05/k where k refers to the number of tests conducted (Bush & Moore, 2012). However, the Bonferroni correction is typically thought too strict, as it assumes each test is independent, that is not the case in GWAS due to LD as discussed above. Other approaches such as the false discovery rate (FDR) procedure developed by Benjamini and Hochberg and permutation testing, are also widely applied in correcting the significance threshold (Bush & Moore, 2012). Currently, a threshold of $5 \times 10^{-8}$ is widely used for the identification of the association between common variants and traits in human studies (Fadista et al., 2016). However, this threshold may vary in other species depending on factors such as genome size and complexity, LD, and population structure.

Calculated p-values can then be presented as Manhattan plots on a genomic scale as shown in Figure 1.4 (B). The X-axis shows the positions of the associated variants along the chromosomes and the corresponding p-values are transformed to $-log_{10}^{p-value}$ for better representation on the y-axis. Variants that are above the threshold for genome-wide significance on this transformed scale can be used to determine the regions of interest for further analysis such as fine-mapping. These regions are typically explored according to the LD structures among the variants and are “zoomed in” for a better illustration.

### 1.1.3.4 Statistical fine-mapping

In the “zoomed in” plots for the regions of interest as shown in Figure 1.4 (B), there is commonly one or several lead variants with the smallest p-value in this region and further multiple significant variants nearby in strong LD. Although the lead variant(s) has the strongest association with the trait of interest in terms of the statistical level, it is not necessarily the precise causal variant. One of the reasons is that the real causal variant may not be genotyped because of the limitations of the genotyping approaches used in GWAS and only this tag SNP, which is in LD with the causal variant, is measured. Furthermore, some other factors may also lead to the uncertainty of the lead variant being the causal one, such as small variant effect size on complex traits (Schaid et al., 2018). These factors emphasize the importance of well-designed fine-mapping approaches for causal variant identification.

There are three main strategies used for fine-mapping: heuristic methods, penalized regression models and Bayesian methods (Schaid et al., 2018). Heuristic methods utilize the LD structure properties in the subregion and measure the correlation among the variants surrounding the lead SNP to determine the potential causal variants. This type of approach has significant limitations as it does not take into account the joint effects of the variants on the phenotype, as well as the lack of objective measurement of the reliability of the results (Schaid et al., 2018). Another type of method uses regression models to analyse variants in the subregions. Traditional regression models use p-values to determine the inclusion of the variants in the models but the stability of the models highly depends on the number of SNPs and their correlations. An advanced approach for dealing with a large number of variants is to use penalized regression models, which reduce the coefficient of SNPs with small effect on the trait in the model down to zero and only keep the subset of variants most associated with the trait (Ayers & Cordell, 2010). Typical penalized models include lasso and minimax concave penalty (MCP). (Schaid et al., 2018)

Another popular strategy used in fine-mapping that can outperform both heuristic methods and penalized regression models is Bayesian approaches (Schaid et al., 2018).

Posterior inclusion probabilities (PIP) of the causality are used in Bayesian approach for quantifying the evidence of a variant being causal (Hutchinson, Watson, et al., 2020). The outputs of Bayesian fine-mapping approach are credible sets which include variants that contain a causal variant with over 95% probability. Each credible set can be viewed as corresponding to one putative causal variant and it is also reflecting the uncertainty around which variant is the actual putative causal variant. As shown in Figure 1.5, the blue credible set only includes one variant, indicating a high level of confidence in this variant being the causal one. In the orange credible set, there are five possible causal variants, suggesting that one of these five variants may be the second causal variant, but it is uncertain.

Bayesian methods can also be used in fine-mapping with the assumption that there are multiple causal variants in the region, which is often the case in many polygenic traits that are affected by multiple variants. There are two approaches for achieving this. One is to split the locus into smaller segments and apply the basic single-causal-variant fine-mapping method to each chunk. Another method is to jointly model multiple causal variants in one Bayesian model for the locus, which is widely employed in many popular Bayesian-based fine-mapping approaches such as CAVIAR, DAP-G and SuRiE. (Schaid et al., 2018).

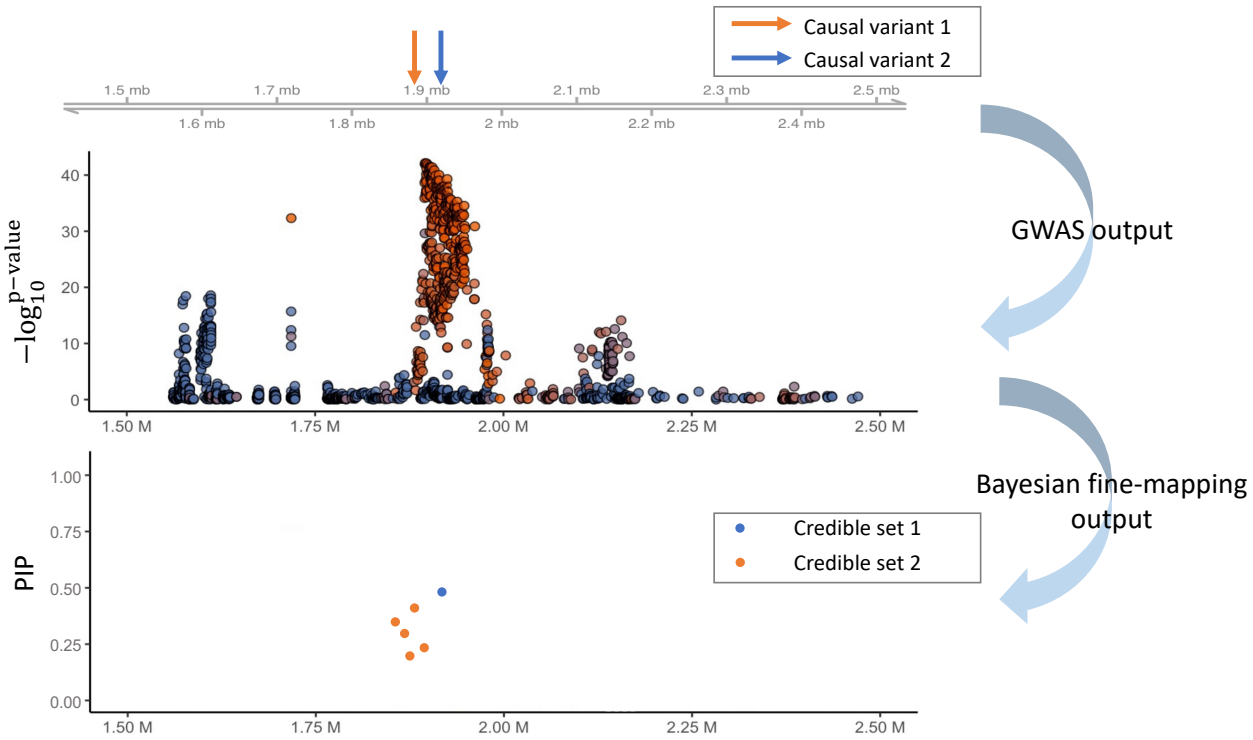



**Figure 1.5 Bayesian fine-mapping outputs.** Bayesian approaches use posterior inclusion probabilities (PIP) to quantify the evidence of a variant being causal and output credible sets according to PIP. The number of credible sets depends on the assumption of numbers of causal variants in the locus of interest.

### 1.1.3.5 Functional inference of the variants

Although fine-mapping has provided sets of reliable causal variants, the underlying biological mechanisms of how these variants impact a downstream phenotype are typically still difficult to infer. According to GWAS, the vast majority of the variants associated with specific traits fall within the non-coding regions of the genome, way more than those residing in protein-coding regions, and this complicates the functional inference of most identified causal variants (Uffelmann et al., 2021).

Genomic annotations that provide biological function insights of the DNA sequences are informative and may be utilized for inferring the possible functions of the variants selected by fine-mapping approaches. There are different types of functional annotations of the variants. The first type is genomic region annotations, such as whether the variant is in a

coding region or not. Chromatin state is also a very important type of annotation as variants in open chromatin regions are more likely to be involved in biological processes (Thurman et al., 2012). Chromatin state annotations can be acquired from DNaseI hypersensitive sites (DHSs). Histone marks, such as methylation and acetylation, are also critical annotations, which provide another source of evidence that a variant plays a functional role in some biological processes such as chromatin remodelling. Besides, variants in regulatory elements such as enhancers, lncRNA can have potential impact on gene regulation and therefore regulatory genomic annotations are crucial for functional inference of non-coding region variants. In general, genomic annotation is a follow-up work of fine-mapping whereas some researchers also leverage these functional annotations to improve the fine-mapping process. For example, PolyFun is a computationally scalable framework utilizing functional annotations across the genome to calculate prior probabilities for fine-mapping approaches and try to improve the accuracy of the identification of causal variants (Weissbrod et al., 2020).

Furthermore, determining the possible affected genes for non-coding causal variants is crucial. Molecular quantitative trait loci analysis (molQTLs) is one of the approaches for the identification of the affected genes of the variants (Uffelmann et al., 2021), which aims to find associations between variants and specific molecular phenotypes. For example, expression quantitative trait loci analysis (eQTL) is one of the molQTLs approach aimed at identifying loci associated with gene expression, while splicing quantitative trait loci analysis (sQTL) focuses on the identification of loci associated with alternative splicing (Kerimov et al., 2021). Projects like the Genotype-Tissue Expression resource (GTEx) catalogues these information in different human tissues (GTEx Consortium, 2020). Recently, similar projects have focused on cataloguing eQTLs and sQTLs in livestock species such as the cattle GTEx program (S. Liu et al., 2022). More details of GTEx and cattle GTEx will be discussed in Chapter 4.

Various tools have been developed for predicting the potential impact of the variants, most of which are based on machine learning methods. These tools utilize large amounts of variant data and their functional annotations as the sample data in machine learning

algorithms to find the potential patterns underlying the dataset that can be used for distinguishing functional variants. This will be discussed in detail in the next section on machine learning.

## 1.2 Machine learning

The concept of machine learning came into the picture in 1950 when Alan Turing, a pioneering computer scientist, proposed the Turing test, which is designed to test whether a machine's ability of exhibiting intelligent behaviors is nearly equivalent to that of a human. Later in 1958, Frank Rosenblot designed the first neural network which is now commonly known as the perceptron model (Rosenblatt, 1958). During the 1990s, work on machine learning gradually shifted from knowledge-driven to data-driven. Researchers began to design machine learning algorithms that can find patterns from large amounts of data and improve their performance upon past experiences with minimal human intervention. In other words, machine learning algorithms are designed for pattern detection within large amounts of data which makes them well-suited to data-driven disciplines such as genomics. This section will first introduce different types of machine learning algorithms and the general processes for machine learning projects, then will delve into the current applications of machine learning in genomics.

### 1.2.1 Machine learning algorithms

According to different learning styles, machine learning algorithms can be divided into supervised, unsupervised, and semi-supervised learning. Supervised learning, as the name suggests, relies on labelled data for training. Labelled data here refers to data with known target answer, such as an image data labelled as a 'dog image', or a variant known to be linked to a disease. The labelled data is fed into algorithms that analyse and learn the associations between data and labels based on features of the data. Features here refer to information acquired according to some properties of the data, such as gender of a human or allele frequency of a variant. Unsupervised learning algorithms are trained using unlabelled data to discover and summarize patterns or structures. Semi-supervised learning stands between supervised and unsupervised learning by utilizing a small

number of labelled data bolstering a larger amount of unlabelled data in the learning process.

#### 1.2.1.1 Supervised learning

Supervised learning can be further divided into two types: classification and regression. The purpose of a classification problem is to assign test data to specific categories, while regression aims to learn the relationship between features and a continuous variable. Below, I discuss some basic example algorithms commonly used in supervised learning.

- **K-nearest Neighbours (KNN)**: This is one of the simplest supervised algorithms that can be used for both classification and regression tasks. The algorithm is based on the assumption that similar cases exist in close proximity (Kramer, 2013). The underlying concept of KNN is relatively simple as it only stores the labelled training data and makes a prediction based on the values or classes of k nearest neighbours of the query point. KNN is easy to implement but becomes increasingly inefficient with large datasets. Furthermore, KNN can be sensitive to outliers and noisy data.

- **Support vector machine (SVM)**: SVM is one of the most robust prediction models which can be used in both classification and regression (Noble, 2006). As shown in Figure 1.6, the basic idea of SVM is to map the original data from low-dimensional space to high-dimensional space in which the data becomes separable and find a hyperplane that can separate between the different classes. The mapping is achieved using kernel functions, including the linear kernel, polynomial kernel, and Radial basis function (RBF). SVM is effective in high-dimensional space but is not suitable for large datasets and may underperform when the number of data features exceeds the number of data.

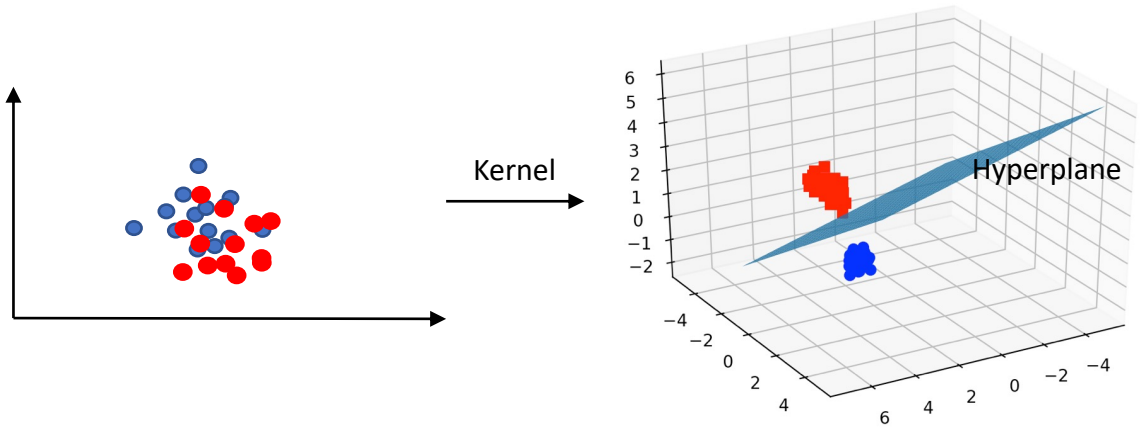



**Figure 1.6 SVM uses a kernel function to map data in low-dimensional space to high-dimensional space within which the data is separable.** The blue plane is the hyperplane that can separate between the two classes.

- **Decision Tree**: Decision Tree is a non-parametric supervised algorithm utilized for classification and regression (Kotsiantis, 2013). It has a tree-shaped structure which includes a root, internal nodes, branches and leaves as shown in Figure 1.7. Each internal node in the tree represents an evaluation to create homogeneous subsets based on a feature, while leaf nodes represent all possible predictions. The root node initially contains all data samples used for model training, and they are subsequently split into sub-groups in the internal nodes based on purity, which means the algorithm is trying to find the ideal attribute to split the data and achieve the highest homogeneity in each sub-group. This "purity" is evaluated by different measurements in different Decision Tree algorithms. The classification and regression tree (CART) uses Gini impurity, which is the probability of misclassifying a datapoint. Iterative Dichotomiser 3 (ID3) tree utilizes entropy and information gain to evaluate splits, where entropy means the measurement of randomness. From root to leaf, a Decision Tree greedily searches for the local optimal at each internal node, iterating this process until it reaches the final global optimal solution.

  Due to their flowchart-like structure, Decision Trees offer the advantage of interpretability and ease of understanding. Besides, they can provide accurate

predictions for large datasets in a relatively short time. However, if there are no restrictions on the splitting depth, the tree will continue to split until the dataset can no longer be split, which will lead to overfitting. Overfitting is a problem that should be considered in many machine learning algorithms. It happens when the model fits the training data too well, including the noise in the data. The overfit model fails to capture the general pattern in the data that can be used for separating different classes and, consequently, performs poorly on new test data. This can be solved by setting limitations on tree depth or cutting some branches from the complete tree. Additionally, in machine learning, various strategies are employed to combine multiple Decision Trees into more powerful models to mitigate the issues associated with individual Decision Trees. Examples of such strategies include Random Forest and XGBoost.

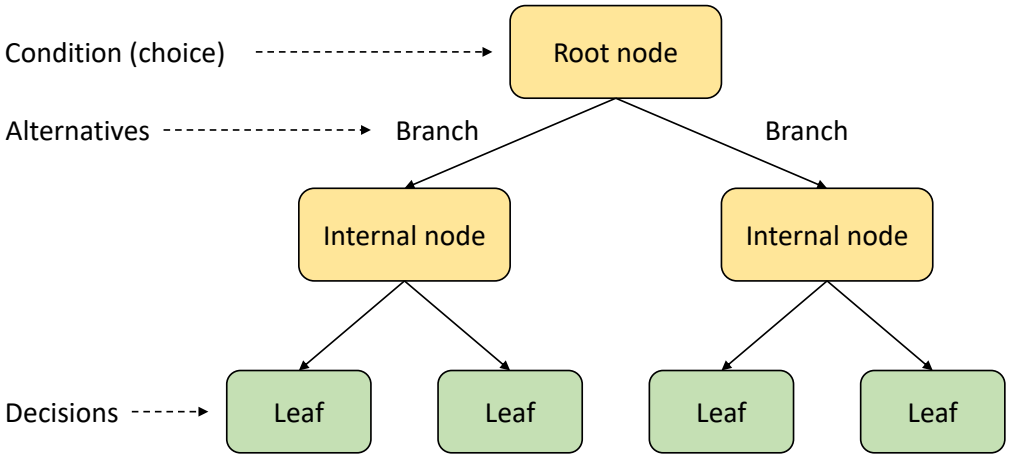



**Figure 1.7 Basic structure of a Decision Tree.** The tree includes a root node, branches, internal nodes, and leaves.

- **Tree models based on ensemble learning approaches:** As discussed in the Decision Tree section, one of the major drawbacks of the most basic tree-based model is overfitting, which hinders the model's ability to generalize well to new data. Therefore, more complex tree-based models have gained popularity in practical applications. These models are usually enhanced by employing ensemble learning approaches, which utilize various strategies to combine multiple weak models and aggregate their predictions to produce final predictions (Dong et al., 2020). One of the

most widely used ensemble learning approaches is Bootstrap Aggregating (bagging). The process of bagging starts with random sampling of the training data with replacement, whereby each subset is used to train an individual base model independently. Subsequently, the final prediction result is determined by aggregating the outputs of each base model through different strategies, such as majority voting (Dong et al., 2020). Random Forest is an application that utilizes the bagging approach to ensemble Decision Trees, which has been widely used in biology, including the study of SNP-SNP interaction and biological sequence analysis (Qi, 2012). Considering the performance and training speed, Random Forest is a good choice to use as a baseline model when handling large dataset with a mix of high-dimensional features of different types.

Another ensemble learning technique is boosting, which differs from bagging in its sequential management of base models. In the boosting strategy, each new individual model is constructed with a focus on correcting the errors made by its predecessors (Dong et al., 2020). Gradient boosting is a specific implementation of the boosting strategy, which employs the gradient descent optimization method to improve the efficacy of the model (Natekin & Knoll, 2013). One of the most powerful ensemble learning algorithm, Extreme Gradient Boosting (XGBoost), is implemented based on the gradient boosting approach (T. Chen & Guestrin, 2016a). XGBoost has been applied to various aspects in biology. For example, it has been demonstrated to outperform models based on neural network or KNN in predicting gene expression values (W. Li et al., 2019). Another merit of XGBoost is its support for GPU during model training and prediction, resulting in a substantial improvement in training efficiency, especially when dealing with large datasets. There are also other tree models based on gradient boosting, including CatBoost (Prokhorenkova et al., 2017) and LightGBM (Ke et al., 2017), each with its own characteristics. CatBoost can not only handle numerical features but also categorical features without beforehand feature encoding. LightGBM is famous for its fast training speed and low memory usage while retaining good performance. Additionally, all these tree-based algorithms offer the capability to obtain feature importance after model training, despite

employing different calculation strategies for the importance scores. This enhances the interpretability of the models for specific tasks.

- **Artificial neural network (ANN)** (Krogh, 2008): The structure of a neural network is inspired by the layered organization of neurons in the human brain. As shown in Figure 1.8, an ANN is comprised of an input layer, multiple hidden layers and an output layer. Each layer includes multiple artificial neurons and each neuron is responsible for part of the computational work. Neurons in one layer are connected to neurons in the next layer through channels and each channel is assigned a numerical value known as a weight. The input of each neuron is multiplied by the respective weight and the sum is then passed through a threshold function called an activation function, which determines whether a particular neuron will get activated based on the comparison with a given threshold. The activated neurons transmit data to the neurons in the next layer and the data is propagated through the network in this manner. In the output layer, the neuron with the highest value is activated and will be used as the final output which is usually a probability of the data belonging to a class. This whole process is called forward propagation. During the training process of ANNs, the predicted output is compared to the desired output and the difference between them is propagated backward through the network to guide the adjustment of the weights. The cycle of forward and back propagation is iteratively performed until a specific stop criterion is met.

  Neural networks are the basis of deep learning, which is a subfield of machine learning. Various types of neural networks are designed to adapt to different data types. For example, convolutional neural networks (CNNs) are commonly applied to image data, while recurrent neural networks (RNNs) are typically used for processing sequential or time series data. Neural networks usually demonstrate good performance, but they are difficult to interpret and computationally expensive due to their complex structure. In addition, training ANNs usually requires large amounts of data.

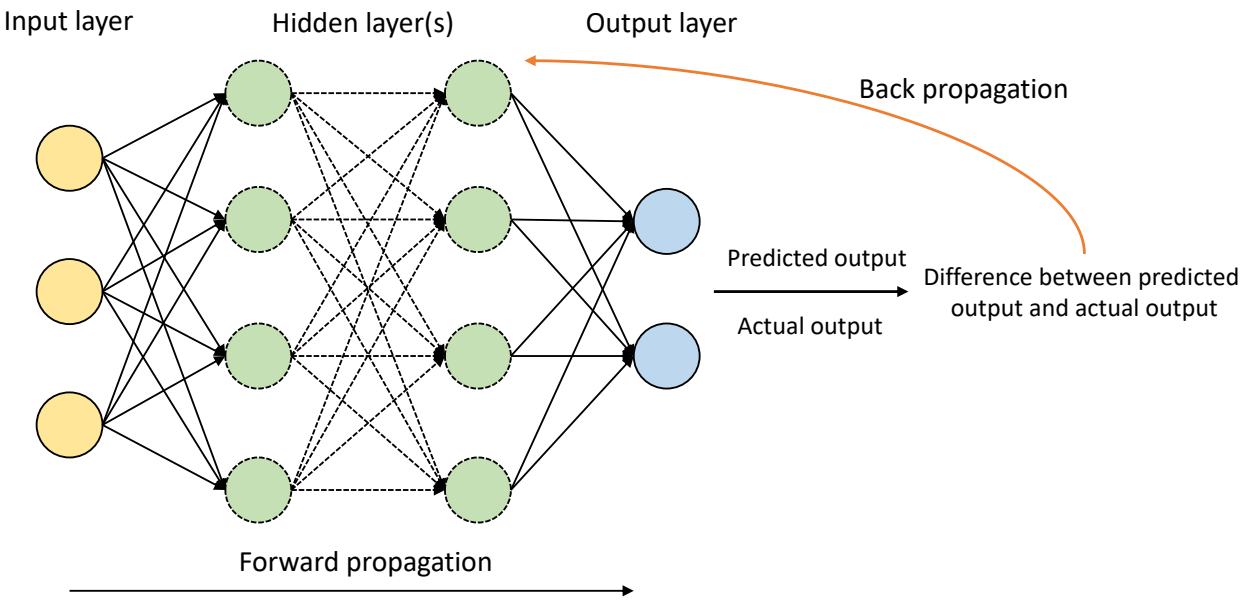



**Figure 1.8 Structure of Artificial Neural Networks.** ANNs include an input layer, multiple hidden layers, and an output layer. Information is passed through the network to obtain the final prediction (forward propagation), and the difference between the predicted and actual output is then used to adjust the weights through back propagation.

### 1.2.1.2 Unsupervised learning

Unsupervised learning uses machine learning algorithms to discover the hidden patterns in unlabelled datasets (Ghahramani, 2004). Clustering, association, and dimensionality reduction are three main tasks in unsupervised learning. The aim of clustering is to group unlabelled raw data based on their similarities or differences, and it is commonly used in applications such as image compression and clustering biological samples. According to different strategies, clustering can be further divided into three categories: exclusive and overlapping clustering, hierarchical clustering, and probabilistic clustering. Association, another method for finding relationships between variables in a dataset, is frequently employed in recommendation systems. Dimensionality reduction is a technique widely used in data preprocessing parts of machine learning projects. Using various strategies, such as principal component analysis (PCA) and singular value decomposition (SVD), dimensionality reduction aims to decrease the number of features in the dataset while

preserving as much of the original information as possible. This can help improve the performance of the machine learning models, as well as their training efficiency.

#### 1.2.1.3 Semi-supervised learning

Semi-supervised learning is in the middle area of supervised and unsupervised learning. In addition to a large amount of unlabelled data, some labelled data is used for supervising the learning process (Zhu & Goldberg, 2009). Semi-supervised learning is suitable for situations where labelled data is insufficient and difficult to get. Self-training is the most basic type of semi-supervised learning. The small amount of labelled data is first used for base model training and then this trained model will be used for pseudo-labelling, i.e., making predictions for unlabelled data. The predictions with high confidence will be added to the labelled dataset and to retrain the model. The whole process usually iterates several times to improve the performance of the model. Co-training is an advanced version of self-training, where two base models are trained using labelled data with different sets of features. In the pseudo-labelling process, each classifier updates its own model using the other's high-confidence predictions, and the predictions from the two updated models are combined to get the final prediction result. Another semi-supervised learning method is graph-based label propagation. In this approach, both labelled and unlabelled data are represented in a graph, and labels are propagated to the unlabelled data points using a label propagation algorithm.

### 1.2.2 Machine learning process

A complete machine learning process typically includes data collection and preparation, feature engineering, model construction and training, model evaluation, hyper-parameter tuning and prediction. Data is at the core of machine learning and determines the upper limit of downstream model performance. Therefore, the initial dataset needs careful preprocessing before being fed into the constructed models. This section will start from feature engineering and then delve into the key steps in the machine learning process.

#### 1.2.2.1 Feature engineering

Most traditional machine learning algorithms take input data in a tabular form. The rows of this form represent the instances and each column refers to a measurable property of the instance which is also known as a feature. For example, when utilizing machine learning in functional variant prediction, each row is a variant and the columns are properties of the variant such as variant position, distance to genomic elements. Therefore, choosing informative and discriminating features is critical in machine learning projects and this highly depends on domain knowledge of the problem being solved. After collecting the raw data and their features, this initial feature table requires some feature engineering steps before it can be used as input in a machine learning algorithm. Feature engineering not only transforms initial features to be compatible with downstream algorithms but also helps improve model performance. Figure 1.9 shows some potential approaches for feature engineering, including data cleaning, feature preprocessing, feature selection and dimensionality reduction. The following section will focus on the most important aspects of feature engineering.

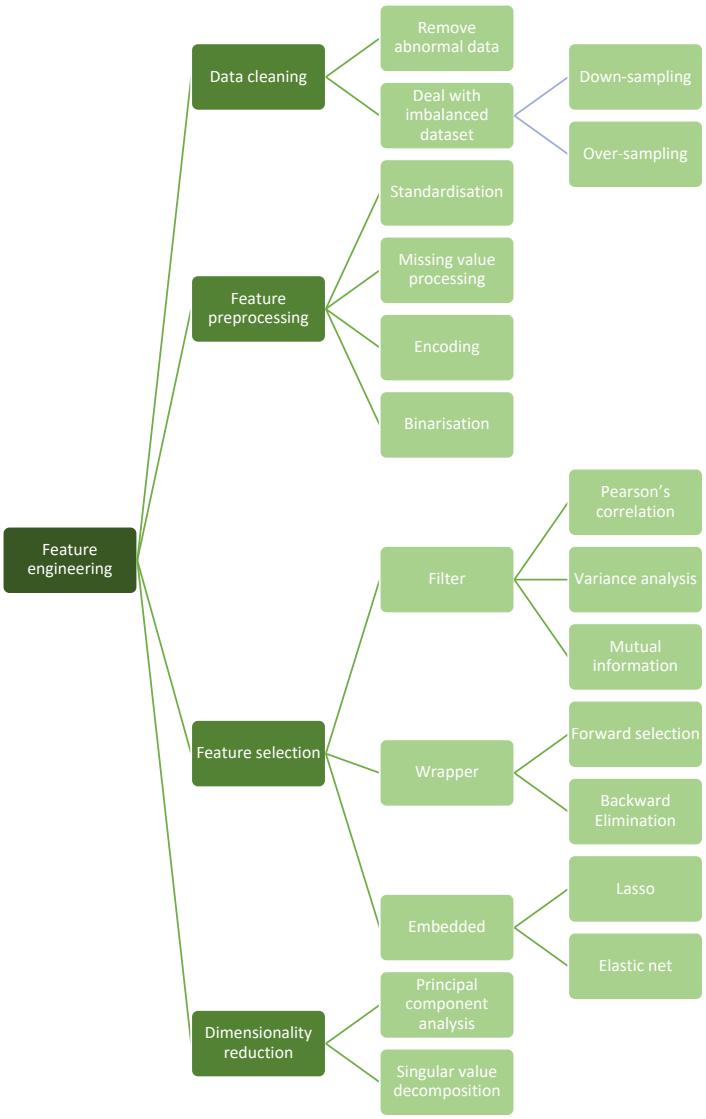



**Figure 1.9 Potential steps and approaches in feature engineering**

Data cleaning is the initial step in the feature engineering process. This includes the process of dealing with the abnormal data that may be incorrect, incomplete, or improperly formatted. Besides, class imbalance is a common problem in machine learning. In binary classification, this problem occurs when one class's proportion is significantly higher than the other class in the dataset. Class imbalance will potentially result in a high prediction accuracy in the majority class but unsatisfactory performance in predicting the minority class. One effective way to solve this is to randomly down-sample the majority class and use the subset to achieve the balance. Another approach is oversampling the

minority class, which involves duplicating instances in the minority class. However, this may lead to overfitting issues within the minority class (Zheng & Casari, 2018).

After data cleaning, the next step is feature preprocessing, which is the most important part in feature engineering. This step usually starts from missing value processing. Many real-world datasets include missing values (often represented as “Nan”) due to a variety of reasons and not all machine learning algorithms can accept Nan values. The simplest strategy is to discard the rows or columns with missing values from the dataset. This is feasible when there is a large amount of data, and the data with missing values accounts for only a small proportion of the whole dataset. However, it is not desirable for small datasets, as it may result in the loss of valuable data. Therefore, missing value imputation is a commonly used method, which includes univariate imputation and multivariate imputation (Zheng & Casari, 2018). Univariate imputation considers only the non-missing values in the same feature column to impute the missing values while multivariate imputation utilizes the entire set of features for imputation. The most widely used method in univariate imputation is to replace the missing values using the mean, median or most frequent values (mode) in the corresponding feature column. For the multivariate imputation, it constructs a function between the feature with missing values and other features to estimate the replacement values.

Another key process in feature preprocessing is to encode the categorical features to make them compatible with machine learning algorithms. Categorical features refer to the qualitative features that can be classified into categories (Zheng & Casari, 2018), such as gender and breed. As most machine learning models can only accept numerical features, it is necessary to transform categorical features into numerical format. According to intrinsic hierarchy, categorical data can be further categorized into nominal data, which does not have an inherent order, such as female or male, and ordinal data, which has a ranking, such as levels of education. There are two commonly used encoding approaches: ordinal encoding and one-hot encoding. Ordinal encoding simply assigns each category in the feature an integer value. This approach is suitable for some ordinal features as the assigned integer values can naturally reflect the ordinal relationship that can be captured

and learnt by machine learning algorithms. However, imposing an ordinal relationship on those nominal features may lead to machine learning algorithms learning misleading patterns, ultimately resulting in poor model performance. The one-hot encoding method is designed for nominal features, replacing them with binary variables instead of integers. One-hot encoding creates binary columns, each indicating the presence of a specific category. This approach informs the model that the original feature is nominal and has no intrinsic order. In addition to these basic encoding methods, categorical features can also be encoded according to self-defined methods that are designed based on the characteristics of the features.

For numerical data, standardisation is often necessary when features in the dataset have different ranges. Standardisation is used to convert numerical features in different ranges to a common scale, without changing the original patterns of the data (Zheng & Casari, 2018). One of the standardisation methods is to standardise the feature by subtracting the mean and scaling the data to unit variance. Another common method is to scale each feature to a given range by its maximum and minimum values. The choice of standardisation method mainly depends on the downstream models used. Binarisation is an approach used for transforming numerical features into binary form. While this transformation is not always necessary, it may enhance the efficiency of the downstream algorithms. In some cases, both the numerical data and the corresponding binary features are retained for feature derivation.

Feature selection is the process of reducing the number of features to an optimal subset which can eliminate non-informative or irrelevant features and thus reduce the training time and improve model performance. There are three main methods for feature selection: filter, wrapper and embedded.

- **Filter**: This kind of method focuses on the features themselves without considering the specific machine learning algorithms that will be used. Features are rated for divergence of correlation with the target and filtered by threshold or the number of features needed. Some common filter methods include Pearson's correlation, variance analysis and mutual information.

- **Wrapper**: These methods initially select a subset of the features, followed by training a machine learning model on the chosen features to evaluate performance. This step is repeated until the best subset of features is selected. The optimal feature subset may vary depending on the specific machine learning algorithm used.
- **Embedded**: The embedded method is a combination of filter and wrapper methods. It relies on penalty terms and is implemented through algorithms that have built-in feature selection mechanisms. An example of this is the LASSO algorithm, which employs L1 regularization for feature selection.

Dimensionality reduction is another method to reduce the dimension of the feature table and reduce training time as mentioned in the unsupervised machine learning section. This kind of method maps the original high-dimensional feature space into a new low-dimensional feature space, where the features in the new space are often combinations of the original features. Principle component analysis (PCA) and linear discriminant analysis (LDA) are two basic methods for dimensionality reduction. The main limitation of dimensionality reduction is the inability to retain the original features, which results in relatively poor interpretability of the new features.

#### 1.2.2.2 Model construction and training

As introduced in Section 1.2.1, machine learning algorithms can be divided into supervised learning, unsupervised learning, and semi-supervised learning according to different learning styles. The choice of machine learning method depends on the problem to be solved. Generally, if the goal of a problem is to make predictions for new data based on the “knowledge” learnt from the known labelled data, supervised learning is the best choice, whereas if the aim is to discover the inherent structure or pattern of unlabelled data, unsupervised learning is more suitable. For example, supervised learning is used for the identification of regulatory elements such as transcription start site (TSS or not TSS) (Libbrecht & Noble, 2015). Unsupervised learning is applied to genomic segmentation and clustering (Hoffman et al., 2012).

The choice of a specific algorithm in the selected learning style is based on various factors such as data type, data volume, computational resources, etc. For structured data, i.e., data presented in a tabular form, traditional machine learning algorithms such as SVM, Decision Tree, Random Forest, should be considered first. However, for unstructured data such as image or text, algorithms with more complex structures like neural network-based deep learning algorithms are more suitable as they can extract features from unstructured raw data. Additionally, data visualization can be helpful in some cases by providing an intuitive presentation of how the data is distributed, thereby guiding the choice of potential algorithms. In practice, determining which algorithm performs better is challenging when relying solely on theory without experimentation. Additionally, one should consider the availability of computing resources, as training models with high-dimensional feature tables or using complex algorithms typically demand significant computational resources.

Machine learning model training is an iterative process that aims to maximize the utilization of the training data. After feature engineering, the entire dataset is initially randomly split into training and test sets. The training set is used for the training process, while the test set is kept separate and only utilized to evaluate the performance of the final trained model. In the training process, k-fold cross-validation is often employed to evaluate model performance on unseen data. K-fold cross-validation is a resampling technique that randomly splits the training set into k folds, with k-1 folds used for training and the remaining fold held for testing. This process repeats until each fold has been used for testing. As a rule of thumb, commonly used values of k are 5 or 10. Models usually perform well when tested on the data used for training, but this does not guarantee the same performance on unseen data. Cross-validation can help assess the model's generalisation and avoid overfitting.

### 1.2.2.3 Model evaluation and hyper-parameter tuning

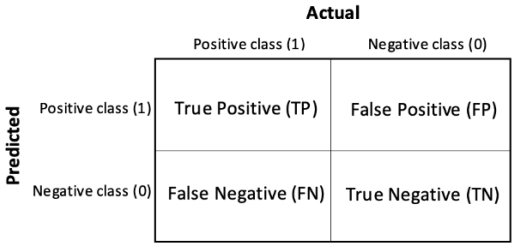
To quantify model performance, different performance metrics can be used. The choice of performance metrices depends on the purpose of the task. For classification tasks, there are three types of commonly used metrics.

- **Accuracy**: Accuracy is the most basic and intuitive metric, and it is defined as:

$$Accuracy = \frac{the\ number\ of\ correct\ predictions}{the\ number\ of\ all\ predictions}$$

Accuracy is a good choice when the dataset is balanced but can become unreliable when evaluating the performance of the model on imbalanced datasets.

- **Precision, recall, F1 score and Matthew correlation coefficient (MCC)**: These metrics are less intuitive than accuracy and are defined based on True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) numbers as shown in the confusion matrix.




TP and TN are correct predictions in both classes while FN and FP are the incorrect predictions that need to be minimized. Precision, recall, F1 score and MCC are defined as:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \quad F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

These equations illustrate the measurements provided by each metric. Precision measures how accurate the positive predictions are, while recall evaluates the proportion of actual positive instances predicted as positive by the model. The F1 score takes both precision and recall into account which is the weighted average of them. The MCC is a more complex metric that considers all the metrics from the confusion matrix. Compared with accuracy, the F1 score and MCC are less intuitive but more useful, especially in imbalanced datasets. Considering a scenario where an imbalanced dataset contains only 10% actual positive instances, the model can achieve a high accuracy of 0.9, even when it predicts all positive instances incorrectly. However, this high accuracy does not actually reflect the model's performance in

predicting positive instances, which is often the primary concern. Therefore, in this situation, F1 score or MCC are more suitable metrics.

- **Receiver Operating Characteristics (ROC) curve and Area Under the Curve (AUC or AUROC):** ROC curves measure the relationship between false positive and true positive rates, with the x axis typically representing the false positive rate, and the y axis representing the true positive rate. AUC is the area under the ROC curve, and it reflects a model's general ability of distinguishing between classes. The higher the AUC, the better the model is at distinguishing between classes.

For regression tasks, several metrics, such as the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R Squared are used. MAE measures the average absolute difference between the actual and predicted values, while RMSE, the most commonly used metric in regression tasks, is similar to MAE but calculates the average of the squared differences. In evaluating the performance of unsupervised learning algorithms like clustering, methods typically rely on measuring the similarity or dissimilarity between clusters. Among these evaluation methods, the Silhouette Coefficient and Dunn's Index (DI) are two of the most employed approaches.

Hyper-parameter tuning is a crucial process for improving model performance. Hyper-parameters are parameters in the model that control the learning process and influence model performance. To further optimize the model, fine-tuning these parameters is necessary. The primary goal of hyper-parameter tuning is to improve the model's performance metric, such as accuracy on unseen data, as seen in Random Forests. Although model tuning is a complex task, the choice of tuning method depends on both the data and the model, there are some basic strategies.

- **Grid search**: This is the most basic tuning method. While tuning, the model is trained and evaluated with each possible combination of the provided hyper-parameter values. Then the best parameter combination is selected according to the performance metric. Although grid search is a powerful method, it is computationally expensive and time-consuming.

- **Random search**: Unlike grid search, the random search method does not require providing discrete values for assessment. Instead, a statistical distribution is specified for each hyper-parameter. In each round of tuning, a parameter is randomly sampled from its distribution, and the performance is evaluated.

### 1.2.3 Machine learning applications in genomics

In recent decades, the amount of biological data, and especially genomic data, has expanded rapidly. As a data-driven approach, machine learning has found extensive applications across various genomic disciplines, ranging from DNA sequence partitioning to gene expression prediction, and has proven to be effective. This section will start by introducing datasets from various fields in genomics that can be utilized in machine learning approaches and then delve into the applications of machine learning in genomics from four aspects.

#### 1.2.3.1 Genomic datasets

Datasets form the foundation of machine learning approaches. To improve the understanding of the subsequent sections, Sections 1.2.3.1 and 1.2.3.2 will introduce the databases/projects or assays for obtaining the datasets involved in the machine learning applications.

- **The Human Gene Mutation Database (HGMD)**: HGMD serves as an extensive repository of published germline mutations in nuclear genes believed to be responsible or associated with human inherited diseases (Stenson et al., 2020). This database currently includes over 289,000 distinct gene lesions found in more than 11,100 genes obtained from over 72,000 peer-reviewed publications.
- **Genome in a Bottle:** Genome in a Bottle consortium, hosted by National Institute of Standards and Technology (NIST), is a collaborative effort for developing the technical infrastructure, including reference data and methods in genomics, to facilitate the integration of whole human genome sequencing into clinical practice (Zook & Salit, 2011).

- **GENCODE:** The GENCODE project generates reference genome annotations for both protein-coding and non-coding loci in human and mouse (Harrow et al., 2012). To date, it has incorporated thousands of annotations for various biotypes of genes, including protein-coding, lncRNA, and pseudogenes, in both the human and mouse genomes.

### 1.2.3.2 High-throughput sequencing techniques

Different experimental approaches are designed to obtain various genomic data, such as regulatory elements and chromatin accessibility. DNase I Hypersensitive Site Sequencing (DNase-seq) utilizes DNase I enzyme to cleave accessible chromatin regions, thereby revealing hypersensitive sites that are linked to regulatory elements (Song & Crawford, 2010). Assay for Transposase-Accessible Chromatin Sequencing (ATAC-seq), an alternative to DNase-seq, utilizes a transposase enzyme to insert adapters into open chromatin regions, enabling the mapping of chromatin accessibility (Buenrostro et al., 2015). Cap Analysis of Gene Expression (CAGE) is a technique for monitoring transcription start sites (TSS) activities by adding the cap structures to the 5' end of RNA molecules, which is effective for promoter and enhancer identification (Morioka et al., 2020). Chromatin Immunoprecipitation Sequencing (ChIP-seq) is a widely used technique for genome-wide DNA binding proteins and histone modifications identification (Park, 2009). Datasets generated from these different techniques provide opportunities to understand functional genomic elements and uncover the intricacies of gene regulation. To date, there have been several studies that have processed and organized various types of genomic data from these techniques. For example, EpiMap is a compendium comprising thousands of epigenomic tracks in humans based on 18 different marks/assays, such as ATAC-seq, DNase-seq and H3K4me1, spanning multiple tissues (Boix et al., 2021).

### 1.2.3.3 Genomic elements identification

Genomic regulatory elements such as enhancers and promoters are responsible for gene regulation and the accurate identification of these elements is the basis of understanding the underlying mechanisms controlling gene expression. Many studies have focused on

using machine learning algorithms to identify regulatory elements. Liu et al. proposed PEDLA, a machine learning-based framework which can be used for the prediction of enhancers. They trained the model with a feature table in 1114 dimensions, including histone modifications, chromatin accessibility, evolutionary conservation, in the H1 embryonic stem cell line. This model manifested a superior performance and achieved 95.7% accuracy and an 81.0% F1-score for 20 independent test datasets (F. Liu et al., 2016). Li et al. developed a package DECRES based on a neural network for the identification of cis-regulatory elements. They annotated 300,000 candidate enhancers and 26,000 candidate promoters using this model and demonstrated that machine learning methods can accurately predict cis-regulatory regions (Y. Li et al., 2018).

Oubounyt et al. utilized the combination of a convolutional neural network and a long short-term memory model to build a promoter predictor called DeePromoter (Oubounyt et al., 2019). They trained the model using promoter sequences as the positive dataset and a self-constructed negative dataset, which utilized subsequences from the positive promoter sequences to retain some conserved parts. This is to force the model to explore more general features that can distinguish between promoter and non-promoter sequences rather than obvious features such as the TATA promoter motif. This strategy improves the generalization of the model on unseen datasets and DeePromoter outperforms other promoter predictors with a Matthew correlation coefficient (MCC) of 0.92 in human and a MCC of 0.87 in mouse. A recent model called BERT-Promoter, which is based on the state-of-the-art natural language processing (NLP) model BERT, not only focused on promoter sequence prediction but also explored the prediction of their activity levels (weak or strong) (Le et al., 2022). Compared to previous promoter predictors, such as iPSW (PseDNC-DL) (Tayara et al., 2020), BERT-Promoter achieved comparable performance in identifying promoters and superior performance in classifying their strength.

Some work has specifically concentrated on identifying transcription start sites (TSS). Grigoriadis et al. proposed DeepTSS, a deep learning-based architecture for distinguishing between TSS-associated Cap Analysis of Gene Expression (CAGE) peaks

and noisy signals (Grigoriadis et al., 2022). They used the DNA sequence centered on the peak position, corresponding CAGE signal values, structural DNA sequence-based features, together with the phyloP evolutionary conservation score as inputs of four distinct convolutional branches for feature extraction and downstream prediction. DeepTSS outperformed other existing TSS predictors, achieving a precision of 98% and sensitivity of 96%.

Other attempts at using machine learning in genomics have concentrated on predicting gene transcripts. Kong et al. proposed the Coding Potential Calculator (CPC), a support vector machine-based classifier trained with six sequence features to classify protein-coding RNAs (Kong et al., 2007). Baek et al. developed LncRNAnet, which is a long non-coding RNA identification model based on the incorporation of recurrent neural networks and convolutional neural networks (Baek et al., 2018). They used an open reading frame (ORF)-based approach for lncRNA identification, which trained the model with candidate transcript sequences and their corresponding identified ORF indicators. To compare the performance of LncRNAnet with other models such as CPC, Baek et al. trained and tested models on the human lncRNA and protein-coding transcript data from GENCODE (Harrow et al., 2012). CPC outperformed LncRNAnet on sensitivity and achieved better results in terms of specificity, accuracy, F1-score and AUC.

Splicing site prediction is another important field where machine learning algorithms have been applied. Jaganathan et al. proposed SpliceAI, a deep neural network designed for predicting splicing junctions from pre-mRNA sequences (Jaganathan et al., 2019). Considering that the distance between splice donors and splice acceptors can be thousands of nucleotides long, SpliceAI used 10kb flanking sequences centered on the splicing donor or acceptor as input and included 32 dilated convolutional layers in the architecture to capture the information that spans long genomic distances. Jaganathan et al. also applied the precise prediction of splicing junctions to predict noncoding genetic variants that affect splicing. Zeng et al. presented Pangolin, another model based on dilated convolutional neural networks for predicting splicing (T. Zeng & Li, 2022). Pangolin

presents improvements over SpliceAI by making the model work in a tissue-specific manner and can predict both the usage and probability of splicing.

#### 1.2.3.4 Variant effect prediction

Predicting which variants are potentially functional is another key application of machine learning in genomics. Ritchie et al. presented a tool called genome-wide annotation of variants (GWAVA), which used a Random Forest algorithm to predict which non-coding variants are linked to disease based on various variant-specific annotations including open chromatin data, TF binding and histone modifications (Ritchie et al., 2014). GWAVA achieved an AUC score of 0.85 when testing on the variations marked as regulatory variants from the Human Gene Mutation database (HGMD). Kircher et al. proposed a Combined Annotation Dependent Depletion (CADD) framework, which integrates diverse variant annotations into a C score (Kircher et al., 2014). They trained a support vector machine with CADD scores to classify pathogenic variants from common variants and found this model outperformed the models trained using other scores, such as conservation scores. Furthermore, Rentzsch et al. introduced splicing scores into CADD to improve the detection of pathogenic variants (Rentzsch et al., 2021). They incorporated splicing scores from MMSplice and SpliceAI as features into CADD and developed CADD-Splice demonstrated superior performance compared to the original CADD model when tested on Multiplexed Functional Assay of Splicing using Sort-seq (MFASS) data, which encompasses variants that influence RNA splicing. Wang et al. introduced the expression modifier score (EMS) for predicting the cis-regulatory effect of the variants in human (Q. S. Wang et al., 2021). They derived EMS by training a Random Forest model with 6121 features based on variant annotations. EMS outperformed previous models such as CADD, etc., in predicting the putative causal whole-blood eQTLs from GTEx.

Instead of using variant annotations as features, Lee et al. trained a support vector machine with 10-mer DNA sequences directly to predict the impact of regulatory variants (Lee et al., 2015). Their model achieved greater accuracy in predicting SNPs associated with DNase I hypersensitivity than previous methods including GWAVA. Similarly, Zhou et al. also utilized regulatory sequence information and developed DeepSEA, a deep

convolutional neural network-based framework, for predicting the functional effects of non-coding variants (Zhou & Troyanskaya, 2015). They divided the genome into 200-bp segments and used 1000-bp sequences centered on these segments as the model input. Performance of DeepSea surpassed that of previous models, such as GWAVA, when testing on HGMD regulatory variants. Later in 2018, Zhou et al. developed another deep learning sequence-based model Expecto that can accurately predict the transcriptional effects of variants in different tissues (Zhou et al., 2018). They utilized their model to prioritize causal variants within all available GWAS loci associated with diseases or traits. The difference between DeepSEA and Expecto is that Expecto focuses on predicting variant effects on gene expression, whereas DeepSEA can predict variant effects that are not significantly associated with gene expression.

Some studies have also focused on the imbalance problem between the amount of data available for functional variants and all variants when predicting functional variants using machine learning models. Schubach et al. presented a method called HyperSMURF to deal with imbalanced data for disease-associated variant prediction (Schubach et al., 2017). They used a sampling method on the imbalanced dataset and trained a Random Forest model on the balanced data. HyperSMURF outperformed other imbalance-unaware methods like DeepSEA on the imbalanced data on Mendelian disease variants. To deal with the scarcity of experimentally annotated data, Jia et al. developed a semi-supervised deep learning method SSL_dnn for predicting functional non-coding variants (Jia et al., 2021). Different from the supervised methods, which use labelled variant datasets to train the model, semi-supervised methods use a small number of labelled variants bolstering a larger amount of unlabelled data. They assessed the performance of their model using the data from human cell lines GM12878, HepG2 and K562 and compared with other supervised models, SSL_dnn showed the higher AUC.

#### 1.2.3.5 Gene expression and regulation prediction

Predicting gene expression and regulation is one of the major applications of machine learning in genomics. Since non-coding functional variants can have an impact on gene expression and affect downstream phenotypes, these gene expression and regulation

predictors are usually further applied to variant effect prediction. Kelley et al. used a deep CNN to implement an architecture for predicting chromatin profiles (Kelley et al., 2016). They trained the model using DNase-seq data to extract signal features. Subsequently, Kelley et al. modified the architecture of Basset by adding multiple densely connected dilated convolution layers to capture information across long distances of the DNA sequences and developed Basenji for gene expression prediction (Kelley et al., 2018). Unlike Basset, Basenji can accept much larger DNA regions as input and it is trained and tested using data from various assays including CAGE, DNase-seq and ChIP-seq.

Avsec et al. proposed another machine learning architecture, called Enformer, that can predict how DNA sequence affects gene expression (Avsec et al., 2021). They utilized a popular neural network architecture in natural language processing which is based on a self-attention mechanism that can capture long-distance information (up to 100kb) from the sequence data. This mechanism helps overcome the limitations of the traditional neural networks used in approaches like Basenji, where distal regulatory elements are ignored in the model. Enformer outperformed Basenji2 on four different assay types: CAGE, transcription factor ChIP-seq, histone modification ChIP-seq, DNase-seq/ATAC-seq. Furthermore, Avsec et al. also demonstrated the potential of utilizing Enformer predictions for predicting variant effects. Chen et al. developed a deep learning model Sei, which can predict 21,907 chromatin profiles across different tissues and cell lines (K. M. Chen et al., 2022). These predicted chromatin profiles are then integrated and applied to develop a global map of regulatory activities, which assigned genomic sequences with 40 different sequence classes including cell type-specific enhancer classes, promoter, CTCF-cohesin. Moreover, these sequence classes can also be used for quantifying the impact (increase, decrease, no change) of the variants on specific regulatory activities.

#### 1.2.3.6 Variant calling

Accurate variant calling from billions of error-prone, short sequence reads is a challenge work and thus researchers have tried to utilize machine learning approaches to improve variant calling from sequence data. DeepVariant is a variant calling technique based on deep neural networks (Poplin et al., 2018). It uses pileup images of reads aligned to the

reference genome around each candidate variant as inputs to train an inception architecture deep learning model. DeepVariant outperformed a commonly used variant calling tool GATK (McKenna et al., 2010) when testing on the Platinum Genome Project NA12878 data. Ramachandran et al. designed another neural network-based architecture called HELLO aiming at solving small variant calling tasks (Ramachandran et al., 2021). Instead of using pileup images, HELLO designed deep neural networks that can accept sequencing data directly. HELLO achieved better false positive and false negative scores in predicting indels compared to DeepVariant, and it outperformed GATK when tested on SNV calling.

Luo et al. proposed Clairvoyante, a CNN-based architecture for multi-task variant calling using single-molecule sequencing (SMS) reads, which can also predict variant type, zygosity, alternative allele and indel length (Luo et al., 2019). Clairvoyante shows a high F1-score on variant calling from both common variant sites and genome-wide. Khazeeva et al. developed DeNovoCNN, a CNN-based approach for de novo mutations calling (Khazeeva et al., 2022). DeNovoCNN uses encoded alignments of sequence reads as input to a CNN. Compared to existing variant calling approaches including GATK and Samtools (H. Li et al., 2009), DeNovoCNN achieved higher recall/sensitivity and precision when trained and tested on the Genome in a Bottle reference dataset (Zook & Salit, 2011). Furthermore, DeNovoCNN demonstrates robustness across various exome sequencing approaches, as indicated by its consistent test performance with different sequencing data.

## 1.3 Conclusion

Functional variants are important drivers of many diseases and traits in mammals. They can have a direct impact on the amino acid code or affect gene regulation to influence downstream phenotypes. Although some studies like GWAS have identified genomic loci that are associated with specific phenotypes, the accurate recognition of actual causal variants in the region of interest remains a challenge due to factors such as linkage disequilibrium. Computational methods such as machine learning approaches can therefore be applied to help improve the detection of the functional variants.

Machine learning is one of the most popular applications in artificial intelligence where the algorithms are trained with known data, learn from experience and make predictions based on learned knowledge. According to different learning styles, machine learning algorithms can be further divided into supervised learning, unsupervised learning and semi-supervised learning. The common first step in a machine learning project is to figure out the problem to be solved, collect relevant data and extract features that are informative. Then the initial features are preprocessed through feature engineering processes to be compatible with the machine learning algorithms of choice. Different machine learning models are constructed based on the purposes of the tasks and the characteristics of the feature tables. The performance of the model can be further improved by hyper-parameter tuning.

To date, machine learning has been widely used in solving problems in genetics and genomics, such as genomic elements identification, variant effect prediction, gene expression and regulation prediction, and variant calling. Most of the machine learning-based models have outperformed previous models based on traditional computational methods. These successful applications have shed lights on the potential of using machine learning in predicting functional variants, though most have been focused on humans.

## 1.4 Aims and objectives

Although a lot of work has been done in predicting functional variants, most of it is restricted to predicting variants directly linked to downstream diseases or traits. Moreover, most studies are based on human data and use a single category of machine learning algorithm to build the prediction model. Therefore, the aim of this study is to develop a more comprehensive model for predicting mammalian functional variants to improve the detection of novel functional variants, especially in less well annotated species such as livestock.

To be specific, my research objectives are:

1. To develop a reusable variant annotation pipeline based on Nextflow and a machine learning pipeline based on scikit-learn. The variant annotation pipeline can annotate variants not only in human but also in other mammalian species according to different properties. The machine learning pipeline includes feature pre-processing, feature selection and model training processes.
2. Apply approaches to predict where human variants have a direct orthologue in a livestock species and explore the characteristics of human variants with livestock orthologues.
3. Demonstrate the utility of the features from the annotation pipeline for predicting mammalian functional regulatory variants using machine learning approaches in both human and cattle.

# Chapter 2 Reusable variant annotation pipeline based on Nextflow

## 2.1 Introduction

### 2.1.1. Variant annotation resources and tools

Variant annotation is a crucial step in genomic sequence analysis as it assigns functional information to DNA variants. The primary aim of variant annotation is to gather sufficient information, both at variant and gene level, to provide insights into the context of the mutation (Hebbar & Sowmya, 2022). This information is valuable for further variant analysis, including machine learning-based approaches. Variant-level annotation encompasses various aspects such as the variant's location, its frequency and expected impact. Additionally, incorporating gene-level data, such as gene function, gene expression pattern, is beneficial in the annotation process, especially for the annotation of challenging variants, including newly discovered ones (Hebbar & Sowmya, 2022).

Variant annotation relies on the utilization of various datasets and databases that contain genomic data and annotations. Ensembl, a genome annotation database developed jointly by the European Bioinformatics Institute and the Wellcome Trust Sanger Institute, is among the most prominent databases, consolidating genomic and biological data for easy access (Cunningham et al., 2022). Ensembl provides a wide range of genomic data including DNA sequences, gene annotations, comparative genomics, variant information and regulation data. Furthermore, Ensembl provides flexible ways for accessing the data, including the use of a REST server interactive access via BioMart, in addition to the traditional FTP site. The UCSC Genome Browser database is another broadly used tool for exploring genomic data (Karolchik et al., 2003). It contains a vast collection of genome assemblies and genomic annotations. Both these databases offer comprehensive collections of genomic data while Ensembl has broader multi-species support and includes a wider range of organisms compared to the UCSC Genome Browser database. Besides, Ensembl integrates more external data sources from diverse projects, thereby expanding the categories of information available within the Ensembl database

(Cunningham et al., 2022; Nassar et al., 2023). In addition to these comprehensive databases, there are some repositories specifically focused on genomic variant information, primarily tailored for humans. For example, the Single Nucleotide Polymorphism Database (dbSNP) is a public repository that contains information about various types of variants including single nucleotide polymorphisms (SNPs), insertions/deletions and structural variations (Sherry et al., 2001); the Genome Aggregation Consortium (gnomAD) provides annotations of a vast collection of variants obtained from diverse populations, placing particular emphasis on allele frequencies and population specific information (S. Chen et al., 2022).

There are various variant annotation tools developed based on these genomic databases, each with different emphases. Wang et.al developed ANNOVAR, which is specifically designed for annotating single nucleotide variants (SNVs) and insertions/deletions, primarily based on the UCSC Genome Browser Database (K. Wang et al., 2010). This tool enables the examination of variants' functional consequences on genes, identification of variants in conserved regions, etc. To be specific, ANNOVAR offers three types of annotations: gene-based, region-based and filter-based. Gene-based annotations provide information about specific variants affecting known genes. Region-based annotations indicate whether variants overlap with specific regions of interest. Filter-based annotations involve the process of comparing and filtering variants against known variant databases. VARIANT, a variant analysis and annotation tool proposed by Medina et.al, can provide information on identified variants, including consequence types and other annotations extracted from various databases such as dbSNP (Medina et al., 2012). In addition to the basic information, VARIANT also generates diverse annotations that encompass details regarding the regulatory roles, such as transcription factors and structural roles, as well as information on the selective pressures acting on the sites impacted by the variant. Another widely used annotation tool is Ensembl Variant Effect Predictor, also called VEP, which is a robust toolset designed for annotating and prioritizing genomic variants in both coding and non-coding genome regions (McLaren et al., 2016). VEP can annotate not only sequence variants including SNVs, insertions, deletions, etc., but also structural variants with more than 50 nucleotides. It provides

comprehensive annotations for the input variants, including detailed information on their effects on transcripts, proteins, and regulatory regions. It also incorporates allele frequencies and phenotype information for those known variants.

Although these annotation tools can provide a wealth of information about the variants, the annotations they provide are more generic and may not be fully tailored for specific downstream analysis such as machine learning. Furthermore, these annotation tools exhibit bias towards humans, underscoring the need for a tool capable of annotating other mammalian species.

### 2.1.2 Bioinformatics workflow management systems and data review guidelines

Variant annotation tools typically require a series of interconnected tasks and analysis steps, often implemented using a combination of different tools and scripts. However, manually managing these processes can be time-consuming and error-prone. To overcome these challenges, it is essential to utilize workflow management systems specifically designed for organizing and executing complex processes involved in the analysis of large-scale biological data. These systems provide structured frameworks that ensure efficient and systematic management of variant annotation processes, facilitating easy reproducibility of these analyses.

Nextflow is an open-source workflow management system that streamlines the development and deployment of computational pipelines that involve handling large amounts of data (Di Tommaso et al., 2017). Nextflow offers a solution with its expressive domain-specific language (DSL) and intuitive dataflow programming model that allows users to code in their preferred programming language and adapt existing scripts to the workflows. Furthermore, Nextflow adopts a container-based approach and abstracts the compute environments from the pipeline logic which ensures the portability of the workflow, providing users with more flexibility in choosing their deployment environment. Additionally, Nextflow incorporates configurable retry logic and continuous checkpoints to enhance the reliability and predictability of workflows. To be specific, Nextflow pipelines

possess the capability to resume execution from the last successfully completed step if an error occurs, minimizing the need for redundant computations. This feature is particularly useful for workflows that involve computationally-intensive and time-consuming tasks. Moreover, Nextflow automatically handles the distribution of data and computation across multiple processes, alleviating users from the burden of implementing explicit parallelization strategies.

Snakemake is another workflow management system that simplifies the development and execution of data analysis workflows (Köster & Rahmann, 2012). It defines workflows as directed acyclic graphs (DAGs) based on a Python-based language, where jobs are represented as nodes and dependencies between the jobs are represented as edges. According to the dependencies, Snakemake can determine the execution order of the jobs automatically. One of the key advantages of Snakemake is its ability to manage data-driven workflows by automatically recognizing alterations in input files and selectively re-executing relevant parts of the workflow to improve efficiency.

Although both Nextflow and Snakemake offer similar advantages, such as parallelization, compatibility, and portability, Nextflow distinguishes itself by providing built-in support for these features, while Snakemake may require more manual configuration. In contrast, Snakemake may need additional steps to attain equivalent functionality. As a result, Nextflow has gained popularity, particularly in large-scale projects, leading to an expanding collection of reusable pipelines and a more active community. For example, nf-core is one of the most comprehensive communities, comprising a collection of Nextflow analysis pipelines (Ewels et al., 2020).

As mentioned before, variant annotation work includes the organization of genomic data from various sources. Therefore, some guidelines could be helpful for reviewing collected data. PRISMA (Preferred reporting items for systematic reviews and meta-analyses) is a comprehensive checklist that ensures transparent and complete reporting of systematic reviews and meta-analyses (Moher et al., 2009). The PRISMA checklist includes 27 items,

such as information sources, search strategy, and selection process, making it a valuable guideline for ensuring quality and comprehensive data review in variant annotation.

### 2.1.3 Objectives

The objective of this chapter was to design and implement a new variant annotation pipeline based on Nextflow, with the goal of creating a reusable workflow suitable for annotating single nucleotide variants in various mammalian species, with a particular focus on use for downstream machine learning projects. This pipeline was designed to incorporate a wide range of annotations including evolutionary conservation scores, variant position properties and variant consequences. The annotations generated by this pipeline serve as valuable resources that can be subsequently utilized in downstream machine learning approaches for functional variant classification.

## 2.2 Materials and methods

In this section, a detailed exposition is provided on the acquisition processes for different categories of variant annotations. Following that, the implementation of the variant annotation pipeline using the Nextflow workflow management system is described.

### 2.2.1 Sequence conservation

Sequence conservation is an important annotation as it provides valuable insights into the functional importance of genetic variants. Variants in conserved regions are more likely to be functional, as by definition if a region is conserved, genetic variants at the locus are not well-tolerated. Indicating that they may be associated with important biological processes such as gene regulation. Conversely, variants occurring in non-conserved regions are less likely to be functional. Therefore, conservation scores may provide useful information when prioritizing functional variants. In this study, two types of conservation scores, phastCons and phyloP, were included. The phastCons score at each base, with values between 0 and 1, is the posterior probability that each base belongs to a conserved element (Siepel et al., 2005). The higher the score, the more conserved the position. The phyloP score reflects the evolutionary conservation at each site, where a

positive value indicates that evolution is slower than expected and the corresponding site is conserved, while a negative value indicates faster evolution (Pollard et al., 2010).

#### 2.2.1.1 Human conservation scores

To obtain human conservation scores, I initially attempted to use the GenomicScores package in Bioconductor. However, this package was found to be highly inefficient. Considering the requirement to annotate large datasets, bigWig files for phastCons100way, phastCons30way, phyloP100way, phyloP30way were downloaded directly from UCSC (http://hgdownload.cse.ucsc.edu/goldenpath) and the package pyBigWig v0.3.17 (Ramírez et al., 2016a) in Python was used to extract the score for each variant from the bigWig files.

#### 2.2.1.2 Mammalian species conservation scores

For other mammalian species, phastCons and phyloP scores were not publicly available. Therefore, the 241-way mammalian alignment from the Zoonomia project was used to calculate the conservation scores for other mammalian species (Armstrong et al., 2020). Cattle was used as an example animal to get the conservation scores.

The 241-way mammalian alignment is a hierarchical alignment format (HAL) multiple alignment including 241 different mammals and was generated using Cactus (Armstrong et al., 2020), a reference-free whole-genome alignment program. This 806 gigabytes alignment file was downloaded from UCSC (https://cgl.gi.ucsc.edu/data/cactus/241-mammalian-2020v2.hal). Figure 2.1 shows the workflow for calculating conservation scores based on the Cactus alignment. The basic idea of getting conservation scores is to compare the cattle-referenced multiple alignment with the model of neutral evolution. To obtain the neutral evolution model, the reference-free HAL format alignment was initially transformed into multiple alignment format (MAF) by chromosome, with the cattle genome serving as the reference genome. This conversion was accomplished using the hal2maf command-line tool from HAL toolkit v2.4.0 (Hickey et al., 2013). Meanwhile, the general feature format (GFF) file for the corresponding cattle genome (Btau_5.0.1, NCBI RefSeq assembly: GCF_000003205.7) included in the Cactus alignment was

downloaded from NCBI (https://www.ncbi.nlm.nih.gov). The chromosome names in the GFF file were changed to match with the chromosome names in the MAF alignment and only the coding sequences (CDS) were extracted from the GFF file. Subsequently, the bed files containing the CDS and the MAF files for each chromosome were used as inputs of the msa_view command-line tool from the PHAST v1.5 package (Hubisz et al., 2011) to extract four-fold degenerate sites (4d sites) in sufficient statistics (ss) format. The 4d sites refer to genomic positions where coding mutations result in synonymous changes, meaning that nucleotide changes at these sites do not lead to changes in the cognate protein sequence and are therefore traditionally used as neutrally evolving sites. The 4d sites from each chromosome were combined using msa_view tool (Hubisz et al., 2011) to obtain a unified file containing all the 4d sites across the genome that can be utilized for generating the neutral model. The phylogenetic tree was obtained from the HAL alignment using halStats tool in HAL v2.2 toolkit (Hickey et al., 2013) and the branch lengths were removed.

Initially, the entire 4d sites file and the phylogenetic tree were used as inputs of the phyloFit tool from PHAST v1.5 (Hubisz et al., 2011) to estimate the neutral model. However, this process exceeded the time limit (28 days) of the university compute server Eddie (*Edinburgh Compute and Data Facility Web Site*, 2021). Therefore, the CDS files for cattle were subjected to a random down-sampling procedure, reducing them to 40% of their original size. Subsequently, the aforementioned processes were repeated to obtain a down-sampled 4d sites file and the neutral model. In the phyloFit command, the default nucleotide substitution model REV, EM algorithm and MED precision were used as parameter settings.

After obtaining the neutral model, the conservation scores were computed using the phastCons and phyloP tools in the PHAST v1.5 toolkit (Hubisz et al., 2011). The phastCons command requires a multiple alignment, phylogenetic models for both conserved and non-conserved regions as inputs, with the option to estimate the conserved model by scaling the non-conserved model. Thus, the neutral model generated by phyloFit was employed as non-conserved model, with the scaling parameter rho set

to the default value of 0.3. The phyloP command requires a multiple alignment and a phylogenetic model as inputs. The chosen method in phyloP for calculating the conservation scores was the likelihood ratio test (LRT), with the mode set to CONACC. In this mode, positive values were employed to denote conservation and negative values were for acceleration. Besides, the wig-scores option was specified to get base-by-base phyloP conservation scores. These parameter choices were inspired by the settings used in calculating the human phastCons100way and phyloP100way. Due to the computational intensity involved in calculating phyloP scores, the MAF files for each chromosome were further split into chunks using maf_parse from PHAST v1.5 (Hubisz et al., 2011). Subsequently, the results obtained for each chunk were aggregated to produce the final conservation file.

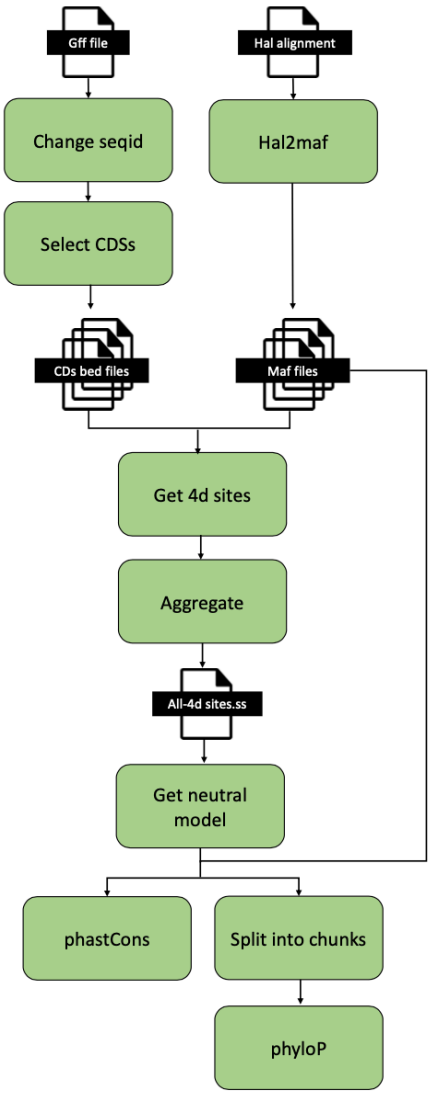


**Figure 2.1 Workflow for getting phastCons and phyloP conservation scores for non-human mammalian species.**

As previously mentioned, the 241-way mammalian alignment utilized the Btau_5.0.1 cattle genome assembly. To obtain conservation scores for the current cattle genome BosTau9 (ARS-UCD1.2), a liftover step is necessary. Although the Cactus alignment does offer a method to update the original alignment file using the new cattle assembly, this process proved to be computationally intensive and exceeded the time limit of our server. Therefore, the decision was made to perform the lift-over step after obtaining the conservation scores for the old cattle genome. As there is no lift-over chain file available for Btau_5.0.1 to BosTau9, nf-LO (Talenti & Prendergast, 2021) was used to produce a chain file. nf-LO is a Nextflow pipeline designed to facilitate lift-overs between any species, provided that their respective assemblies are available (Talenti & Prendergast, 2021). The source and target NCBI accession, GCF_000003205.7 and GCF_002263795.1, were provided to nf-LO. Since this is a lift-over between the same species, the aligner BLAT was utilized and the distance parameter was set to “near” in the nf-LO command. Subsequently, the lift-over was performed by chromosome using the UCSC liftOver tool (Hinrichs et al., 2006).

### 2.2.2 Variant position properties

By determining the proximity of variants to important genomic elements such as genes and regulatory elements, researchers can assess the potential impact of variants on gene function, protein structure and various biological processes more accurately. In order to fully capture the position characteristics of the variant, the distances between the variants and different genomic elements were calculated.

The genomic location of CpG islands of different species was downloaded from UCSC and 1554 different types of human chromatin data from various tissues or cell types were obtained from the Ensembl FTP server (ftp://ftp.ensembl.org/pub). Distances to the nearest CpG island and chromatin data were calculated using bedtools v2.30.0’s “closest” function (Quinlan & Hall, 2010). Transcription start sites (TSS) of different species were obtained using the AnnotationHub v3.2.2 library from the Bioconductor project

(Gentleman et al., 2004) in R and only those common TSS with biotype counts greater than 1000 were retained for further calculation. The human regulatory features, including enhancers, transcription factor (TF), CTCF binding site, were obtained from Ensembl using the biomaRt v2.50.3 package in R (Smedley et al., 2009). The annotatePeakInBatch function from ChIPpeakAnno v3.28.1 library in R was utilized to calculate the distances to the nearest TSS and regulatory features. It should be noted that the distances to chromatin data and regulatory features are specific to humans, as various chromatin data and regulatory features are only available for human from commonly used databases. Gene density per mega base pairs (1Mbp) for different species was also calculated using the AnnotationHub library (Gentleman et al., 2004) and GenomicRanges library (Lawrence et al., 2013). A 1Mbp window was defined for each variant and then the number of genes within this region was calculated using the findOverlaps function in the GenomicRanges v1.46.1 library (Lawrence et al., 2013).

### 2.2.3 VEP annotations

As mentioned in 2.1.1, VEP is a widely used variant annotation tool (McLaren et al., 2016). The VEP v105.0 command line was used to obtain two important variant annotations: variant consequence for different species and allele frequency for human. Variant consequence refers to the predicted functional impact of a genetic variant on the gene or the downstream protein. VEP utilizes a rule-based method to predict the impact of each variant and assigned a Sequence Ontology (SO) term (Eilbeck et al., 2005) to the predicted consequence. Additionally, VEP provides an impact rating for each consequence. Allele frequency denotes the frequency or prevalence at which a particular variant occurs within a given population. VEP provides allele frequencies from different populations (combined population, African, American, East Asian, European, South Asian) sourced from 1000 Genomes phase 3 (The 1000 Genomes Project Consortium et al., 2015), thus making them exclusive to humans.

### 2.2.4 Sequence context

Sequence context of a variant refers to the DNA sequence surrounding the variant, which can provide information about the potential impact of the variant as specific nucleotide

combinations may be more likely to affect gene expression or other regulatory activities. In addition to the basic allele change of a variant, the 5-mer flanking sequence centered on the target variant was also included in the set of annotations. A length of five was chosen to balance sequence information content and computational efficiency in downstream applications. To get the 5-mer flanking sequence, the reference genomes for different species in fasta format were obtained from UCSC (https://hgdownload.soe.ucsc.edu/goldenPath). Samtools v1.12 (H. Li et al., 2009) "faidx" command, which is designed for querying regions from a fasta file, was then used to extract the flanking sequence around each variant from the reference file.

### 2.2.5 Predicted functional genomic data based on Enformer

In addition to these annotations generated using conventional tools or databases, functional genomic data predicted by Enformer was also incorporated as a set of complementary annotations. As introduced in 1.2.2.3, Enformer is a deep learning architecture designed for predicting how DNA sequence influences gene expression (Avsec et al., 2021). The Enformer model was trained using human and mouse genomic intervals (196,608 bp). Each human example within the training set contained a comprehensive set of 5313 genomic tracks, including transcription factor chromatin immunoprecipitation and sequencing (TF ChIP-seq), histone modification ChIP-seq, DNase-seq, ATAC-seq and CAGE tracks. Each mouse genome contained a collection of 1643 genomic tracks. In this study, the pre-trained human model was utilized to predict 5313 different genomic data associated with gene expression and chromatin states for each variant. Moreover, because gene expression patterns and transcription factor binding preferences exhibit significant conservation across mammalian species (L. Chen et al., 2018), the pre-trained human model was also leveraged to other mammalian species to explore the possibility of utilizing the human-based model in other species.

The machine learning library Tensorflow v2.4.1 in Python was utilized and GPUs on Eddie (Edinburgh Compute and Data Facility Web Site, 2021) were employed for efficiency. The pre-trained model was loaded from Tensorflow hub (TF-Hub) (https://tfhub.dev/deepmind/enformer/1) and the reference genome fasta files for different

mammalian species were obtained from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath). A Python package pyfaidx v0.6.4 (Shirley et al., 2015) was used to index the reference genome files. The python code for running the pre-trained model was adopted from the Enformer GitHub page (https://github.com/deepmind/deepmind-research/tree/master/enformer). The pre-trained model loaded from TF-Hub required an input sequence length of 393,216 bp which then got cropped to 196,608 bp centered on the target variant in the model. Hence, the 393,216 bp sequences centered on both the reference and alternative alleles of the target variant were extracted from the reference genome and encoded using the one-hot encoding method, where the order of indices corresponds to 'ACGT'. The model made predictions for the central 114,688 bp of the input sequence, using 128 bp as a unit. Therefore, the prediction output for each input sequence was a 896×5313 matrix. The prediction results for both reference and alternative sequences were obtained, and the score for each variant was defined based on the differences between the reference and alternative prediction results. As a result, each variant was defined as a 1×5313 vector, where each column represents the difference of the predicted genomic data between the reference and alternative sequences.

### 2.2.6 Variant annotation pipeline structure based on Nextflow

To facilitate the reusability of the aforementioned annotation approaches, a SNV/SNP annotation pipeline was developed based on Nextflow. Figure 2.2 shows the basic structure of the annotation pipeline. The pipeline can accept variant files in common formats, such as VCF and BCF, as input. Additionally, it can also accept a 0-based BED-like format with four columns, including chromosome, start position, end position, and variant id following the style of "chr1_10000_C_T", where "chr1" is the chromosome name, "10000" is the variant position, "C" represents the reference allele and "T" represents the alternative allele. The input file in VCF or BCF format is first converted to the 0-based BED format in the pipeline since most annotation approaches require an input file in the 0-based BED format. Meanwhile, according to the species and genome assembly version specified by users, the pipeline generates corresponding species parameters used in different annotation approaches for some common mammalian species including human,

cattle, pig and sheep. The BED-like file is then split into chunks to enable parallel annotation for large datasets, thereby improving annotation efficiency.

After the pre-processing steps, the dataset was then fed into the annotation processes. The aforementioned annotation approaches were encapsulated into functions and scripts and called in the corresponding processes. For variant position annotations, an additional process for calculating the distances from the variant to other genome elements was also included, enabling users to utilize customized data. Due to the parallelization property of Nextflow, these processes can run in parallel according to the computational resources available in the execution environment. It should be noted that the conservation annotation for mammalian species other than human and cattle, as well as the Enformer annotations, were built as two separate and optional sub-workflows due to their high computational time and resource requirements. After obtaining the annotation results for all chunks of the input data, the pipeline collected and organized these results to generate the final annotation results.

The default values of the parameters and profiles for different execution environments, including local, Eddie, were defined in the nextflow.config. The detailed settings for each execution environment, such as task memory, task CPU, maximum memory, were defined separately in the sub-configuration files. To manage the libraries and packages in the environment, the package management system Conda (*Anaconda Software Distribution*, 2021) was utilized, and a YAML file was generated that included the required packages and channels. The source code for the pipeline can be found here: https://doi.org/10.7488/ds/7701.

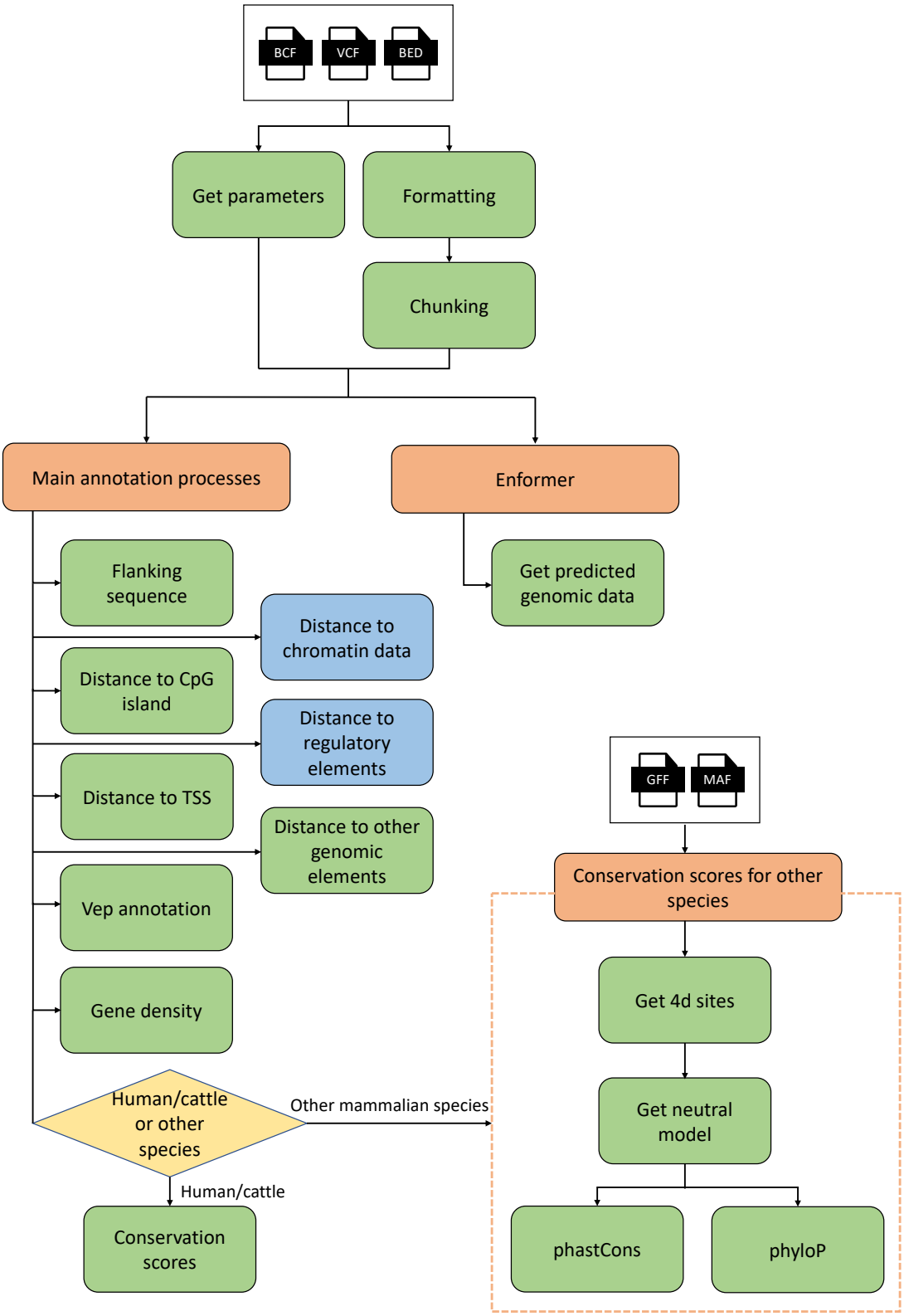

BCF
VCF
BED
Get parameters
Formatting
Chunking
Main annotation processes
Enformer
Flanking sequence
Get predicted genomic data
Distance to chromatin data
Distance to CpG island
Distance to regulatory elements
Distance to TSS
Distance to other genomic elements
Vep annotation
Gene density
GFF
MAF
Conservation scores for other species
Get 4d sites
Get neutral model
phastCons
phyloP
Human/cattle or other species
Other mammalian species
Human/cattle
Conservation scores

**Figure 2.2 Basic structure of the variant annotation pipeline based on Nextflow.** It includes the main annotation workflow and two optional workflows for obtaining Enformer scores and conservation scores for mammalian species other than cattle and human. The two processes in blue are specific to humans.

**Table 2.1 Summary of variant annotations**

| | Annotation | Data source | Main package | Number of features |
|---|---|---|---|---|
| Sequence conservation | phastCons | Human: UCSC<br>Other: Cactus, NCBI | Human: pyBigWig<br>Other: Cactus, PhastCons | Human: 2, other: 1 |
| | phyloP | | | Human: 2, other: 1 |
| Variant position properties | Distance to CpG island | UCSC | bedtools | 1 |
| | Distance to chromatin data | Ensembl | bedtools | 1554 (Human only) |
| | Distance to TSS | Ensembl | ChIPpeakAnno | Human: 14, other: 3 |
| | Distance to regulatory features | Ensembl | ChIPpeakAnno | 4 (Human only) |
| | Chromosome | - | - | 1 |
| | Variant position | - | - | 1 |
| | Gene density (per mega base) | Ensembl | GenomicRanges | 1 |
| VEP annotation | Consequence | Ensembl | VEP | 1 |
| | Allele frequency | | | 6 (Human only) |
| Sequence context | Allele change | - | - | 1 |
| | 5-mer flanking sequence | UCSC | samtools | 1 |
| Enformer | Predicted functional genomic score | Enformer | Tensorflow | 5313 |

# 2.3 Results

## 2.3.1 Annotation pipeline performance

To test the performance of the pipeline at annotating human variants, 1000 and 10,000 variants were randomly sampled from the genome-wide set of 78 million human variants in the 1000 genomes project (Auton et al., 2015). For better clarity, the performance of the main annotation workflow and other two sub-workflows are presented separately. These annotation experiments were conducted on the University of Edinburgh research compute cluster, Eddie (*Edinburgh Compute and Data Facility Web Site*, 2021). The

experiments were performed using different numbers of cores and memory allocations to assess performance differences.

To evaluate the parallelization performance of the main annotation workflow, the set of 1000 variants was annotated using 4 cores with 16GB per core, as well as a single 64GB core. The chunk size was set to 200 variants. As shown in Figure 2.3 and Figure 2.4, each annotation process was executed 5 times due to the input dataset being split into 5 chunks. The most time-consuming process is the calculation of the distance to chromatin data, primarily because of the inclusion of a large number (1544) of chromatin datasets. The processes in the job utilizing 4 cores with 16GB each can run in parallel, allowing the entire job to be completed in an elapsed time of 27 minutes and 44 seconds, where elapsed time denotes the duration from the starting to the ending point of a job. In contrast, the processes in the job utilizing 1 core with 64GB can only run sequentially, resulting in a total elapsed time of 1 hour, 25 minutes and 11 seconds, which is substantially longer than the parallelized job. Furthermore, the execution times for a specific process were found to be similar when comparing sequential and parallel execution, as illustrated in Figure 2.3, Figure 2.4, and Figure S2.1. This is mainly because, as indicated by the memory usage shown in these figures, it is evident that the processes do not require a large amount of memory, with the maximum usage of 3.6GB for the VEP process, which is lower than the allocated 16GB. Thus, the increased memory capacity of the 64GB core may not therefore have a substantial impact on execution time.

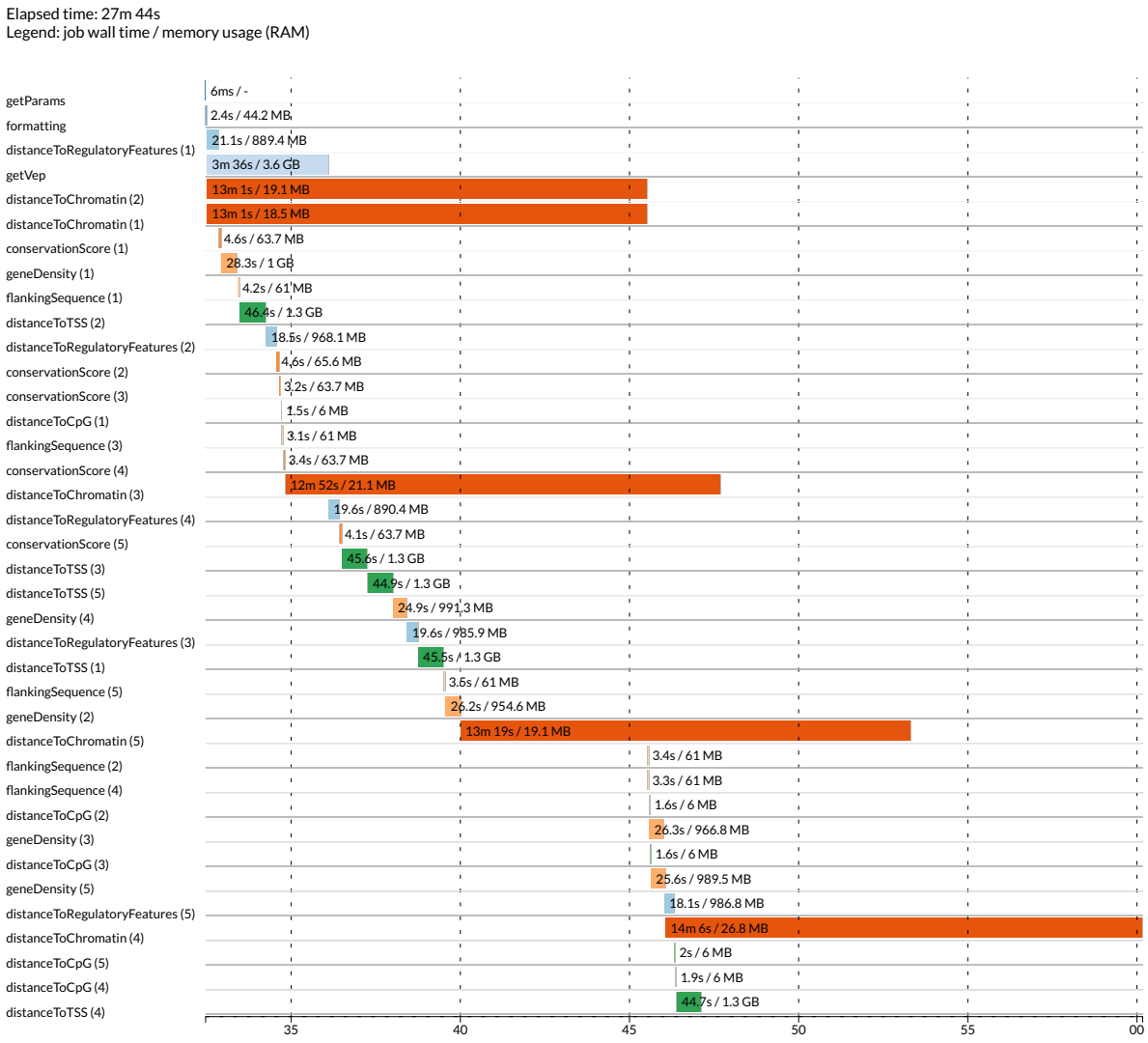



**Figure 2.3 The timeline for all processes executed in the main annotation workflow when annotating 1000 variants using 4 cores with 16GB per core.** The numbers on the x-axis represent time in absolute units. The y axis shows the process names along with their corresponding chunk numbers (ranging from 1 to 5, given that the chunk size is set to 200). Each bar in the plot represents a process and the length of the bar represents the duration time. The numbers on each bar represent the job wall time and the memory size peak respectively, where wall time refers to the elapsed time of a process. The processes run in parallel have the same start point.

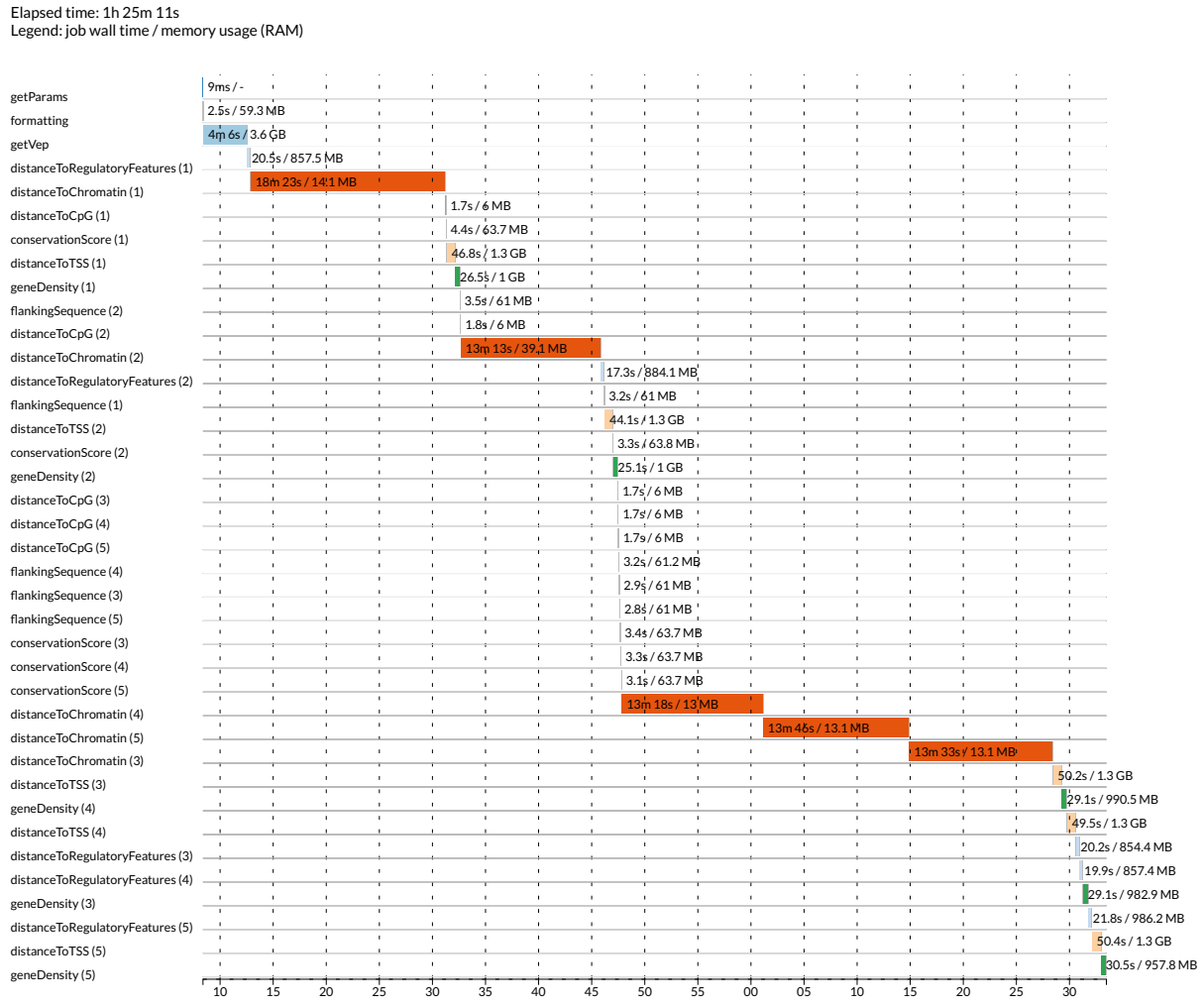



**Figure 2.4 The timeline for all processes executed in the main annotation workflow when annotating 1000 variants using a single 64GB core.**

Subsequently, the set of 10,000 variants were annotated on Eddie using 10 cores with 16GB per core. Different chunk sizes were employed to assess their impact on the execution time. As shown in Figure 2.5, the elapsed time initially decreased with the increase in chunk size, reaching its lowest of 1000 seconds when a chunk size of 2000 was used. However, the elapsed time gradually increased as the chunk size was further increased. This indicates that a smaller chunk size does not necessarily result in a shorter execution time. This is because using a smaller chunk size leads to an increased number of processes, while the limited number of available cores in the environment restricts the number of processes that can run in parallel. The increased number of execution rounds

may lead to an increase in the overall running time. Therefore, the selection of the chunk size needs to take into consideration both the number of available cores for parallel execution and the total number of variants. Figure 2.6 and Figure 2.7 shows the execution timeline for annotating 10,000 variants using chunk sizes of 1000 and 2000 respectively.

Moreover, among multiple experiments, the best elapsed time for annotating 1000 variants using 10 cores with 16GB per core was 15 minutes and 46 seconds (Figure S2.2). A comparable elapsed time of 16 minutes and 43 seconds (Figure 2.7) was achieved when annotating 10,000 variants with a chunk size of 2000 using the same computational resources. This slight time difference observed in annotating a dataset that is ten times larger illustrates that the chunking process, together with the parallelization property, greatly improves the annotation efficiency, making it advantageous for annotating large datasets.

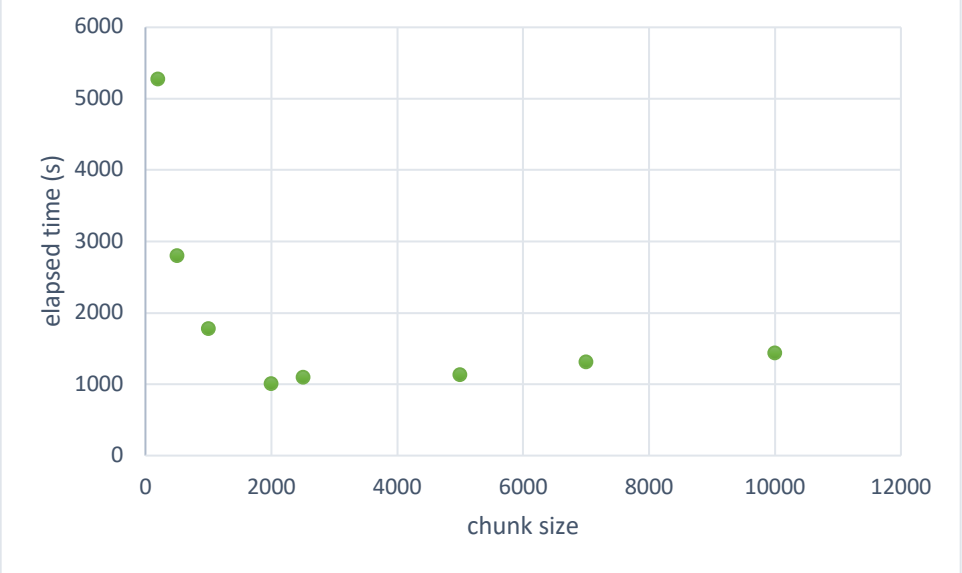



**Figure 2.5 Different chunk sizes and their corresponding job elapsed times when annotating 10,000 variants using 10 cores with 16GB per core.**

Elapsed time: 29m 38s
Legend: job wall time / memory usage (RAM)

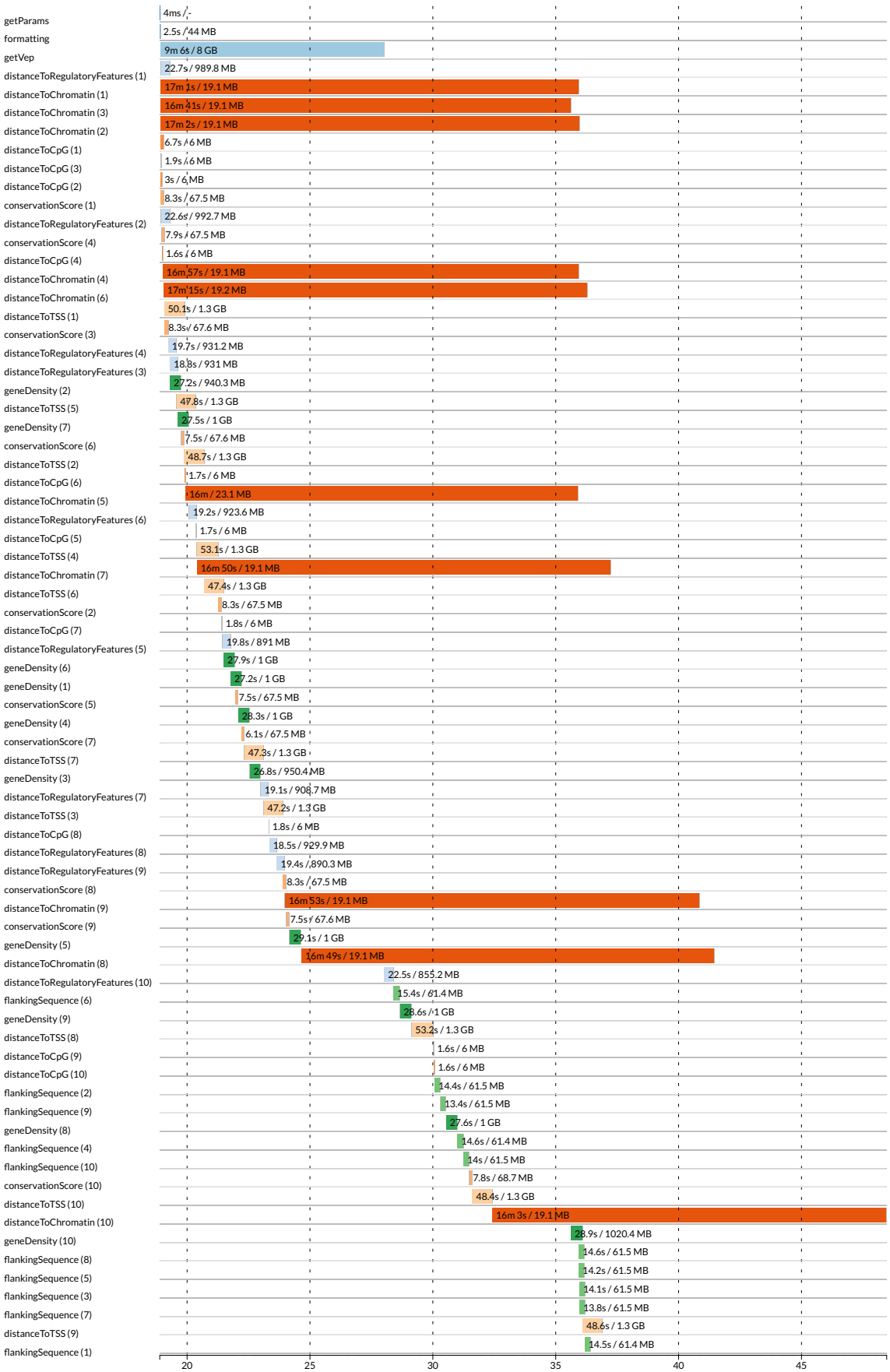


**Figure 2.6 The timeline for all processes executed in the main annotation workflow when annotating 10,000 variants using a chunk size of 1000.**

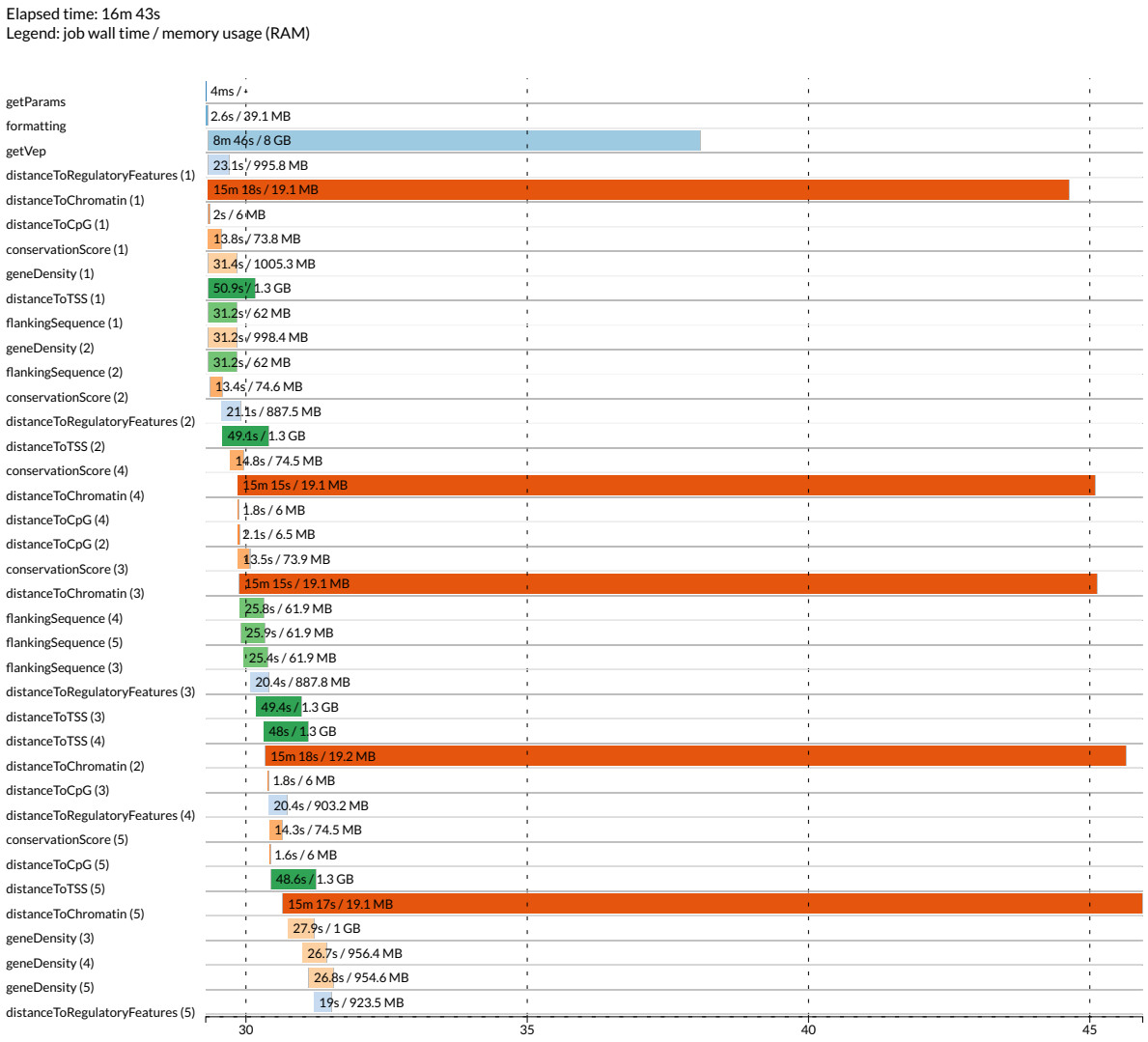



**Figure 2.7 The timeline for all processes executed in the main annotation workflow when annotating 10,000 variants using a chunk size of 2000.**

Additionally, different number of cores were utilized to annotate 10,000 variants with a chunk size of 1000 in order to compare the performance differences of the pipeline under different execution environments. As depicted in Figure 2.8, the job elapsed time exhibited a substantial decrease from the job with 5 cores to the job with 15 cores. Subsequently, it remained relatively stable at approximately 1000 seconds for the jobs with 20 and 25 cores. However, the elapsed time gradually increased when 30 and 35

cores were used. This may be attributed to the following reasons. First, running more processes in parallel can introduce additional overhead and coordination, as tasks need to be distributed and managed across the cores, which can cause an increase in the overall execution time. Second, resource contention such as input/output (I/O) can also result in slower execution. Third, this may also result from the contention for shared resources among the processes. Therefore, it is important to find an optimal balance that maximizes resource utilization and results in better performance.

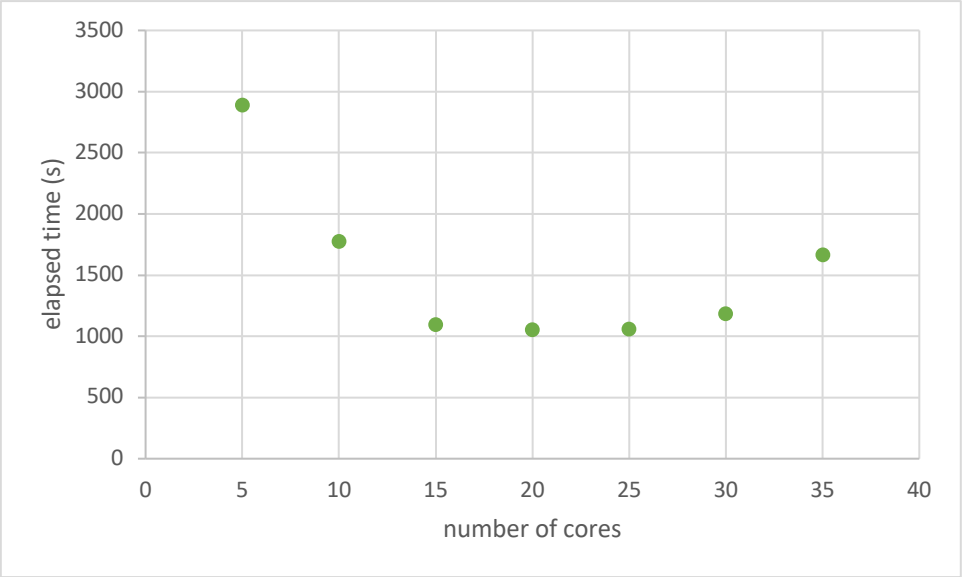



**Figure 2.8 Number of cores (16GB per core) used and their corresponding elapsed times when annotating 10,000 variants with a chunk size of 1000**

To test the performance of the main workflow on annotating other mammalian species, 10,000 cattle variants were randomly sampled from a set of variants generated by Dutta et al (Dutta et al., 2020). The job was executed on Eddie using 10 cores with 16GB per core and the chunk size was set to 2000. Due to the different formats of conservation score source files, the process of getting conservation scores from the source files for cattle is slower compared to humans, and it is the most time-consuming process in cattle annotation as shown in Figure 2.9.

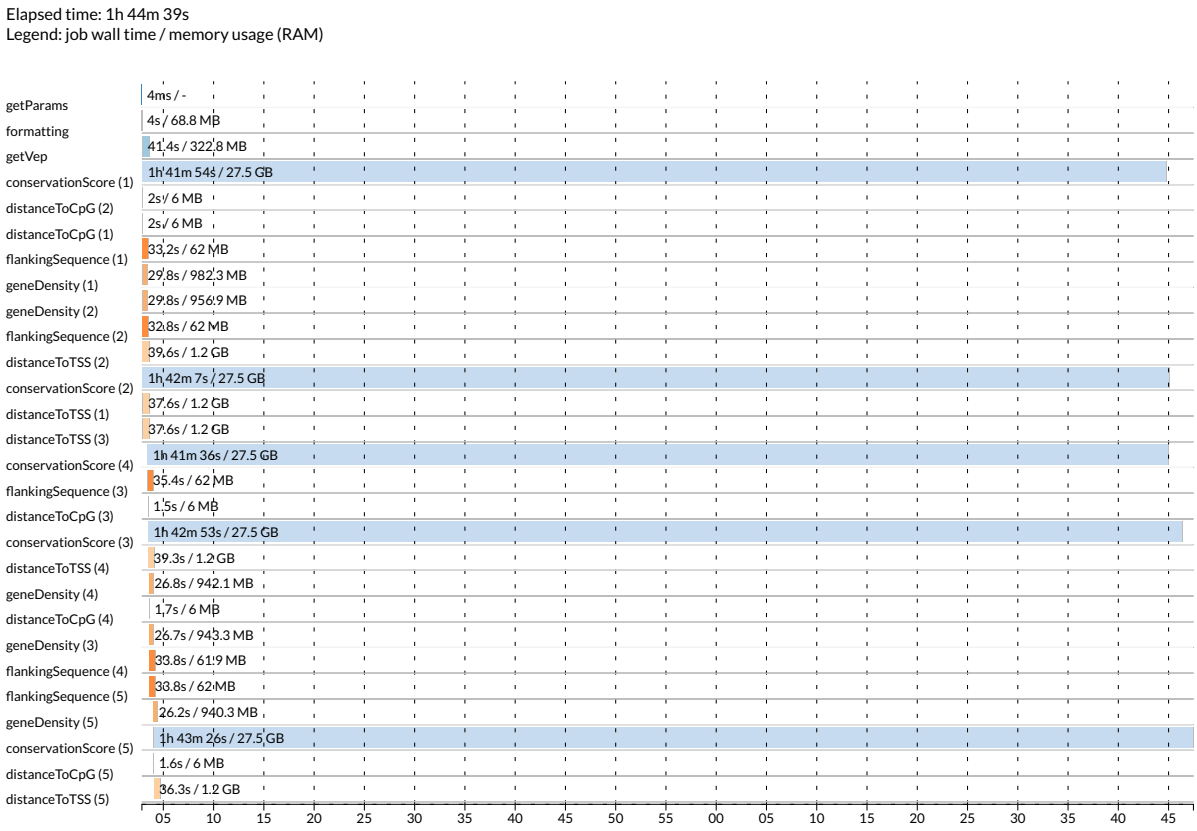



**Figure 2.9 The timeline for all processes executed in the main annotation workflow when annotating 10,000 cattle variants using 10 cores with 16GB per core.**

The performance of the Enformer sub-workflow was tested on Eddie using 2 GPU cores (NVIDIA A100) with 64 GB per core. As shown in Figure 2.10, it took 33 minutes and 12 seconds to obtain predictions for 5313 Enformer regulatory features for 1000 variants. The Enformer sub-workflow can also run on CPUs but it requires much more time (10 hours and 18 minutes when using 10 CPU cores with 16GB per core) to generate the predictions for the same set of variants compared to the GPU environment.

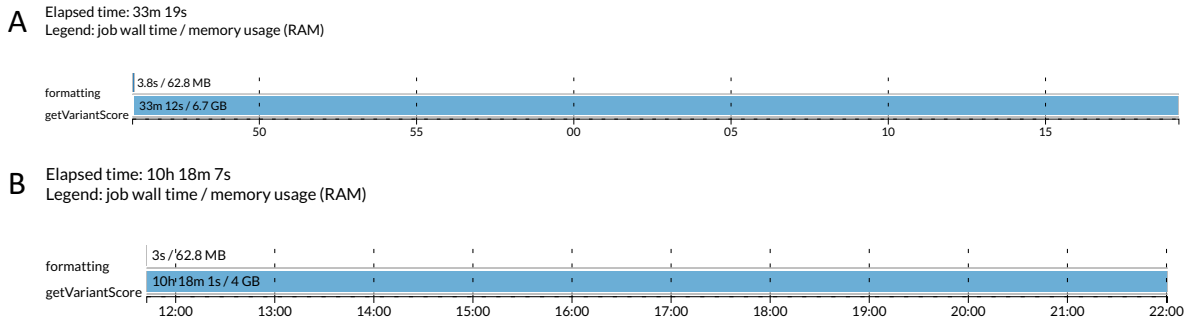


**Figure 2.10 The execution timelines for the processes in the Enformer sub-workflow when annotating 1000 human variants using 2 GPU cores with 64GB per core (A) and 10 CPU cores with 16 GB per core (B).**

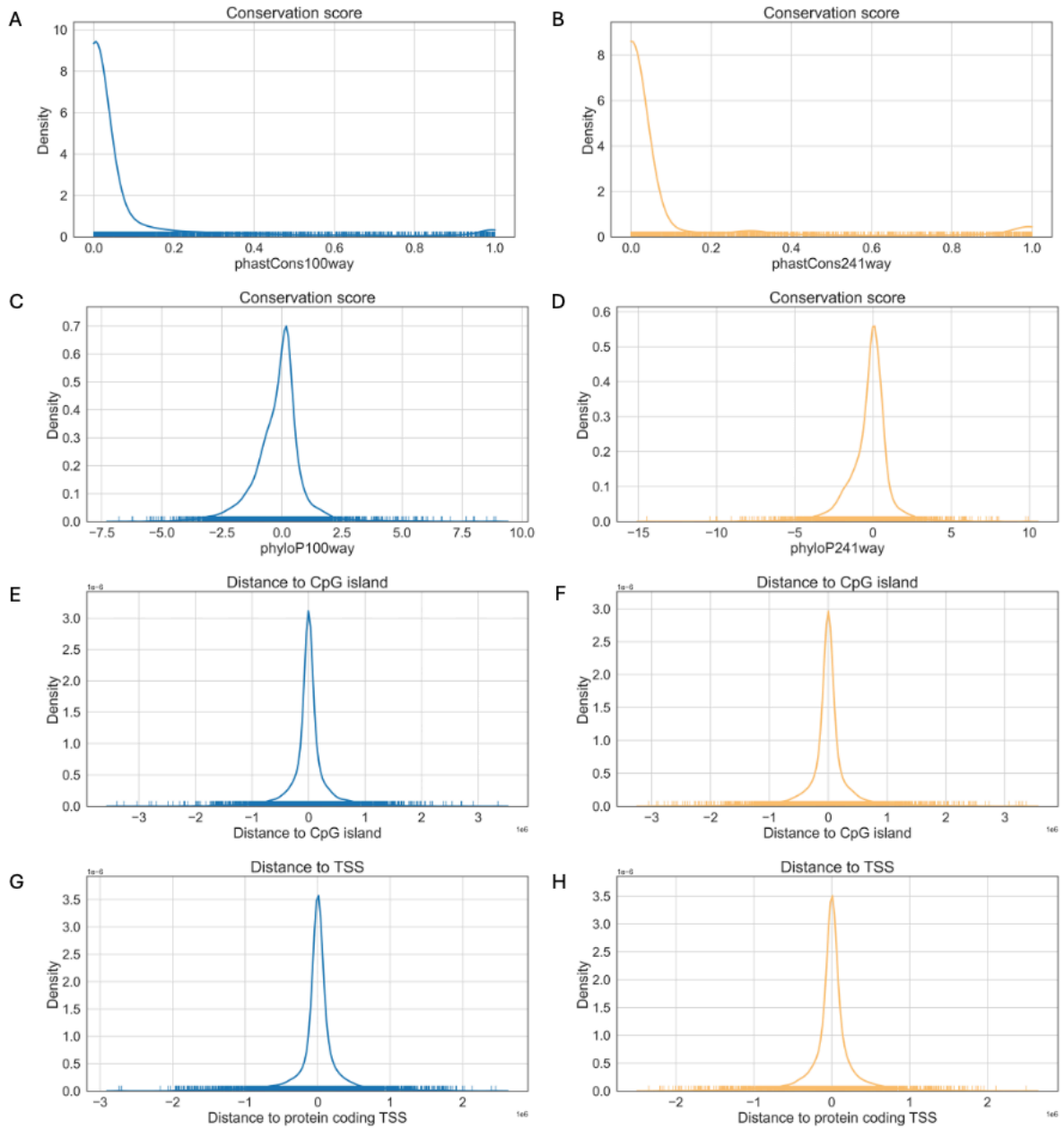
The sub-workflow for calculating conservation scores for mammalian species, except for humans, was utilized for obtaining the cattle conservation scores. Initially, the job failed at the step of estimating the neutral model due to server time constraints when using the 4d sites file (144MB) generated from the whole GFF file. Subsequently, the sub-workflow was executed using the down-sampled GFF file (67MB) and it took approximately 24 days to complete when using 20 CPU cores with 16GB per core. It is highly recommended to down-sample the GFF file to improve efficiency if needed to obtain conservation scores for other mammalian species.

### 2.3.2 Annotation results

To further present the annotation results, distribution plots were made for some of the annotations for the set of 10,000 human variants (Figure 2.11 A, C, E, G, I, K, M, N) and the set of 10,000 cattle variants (Figure 2.11 B, D, F, H, J, L). By comparing the distributions of the annotations, it can be observed that the annotations for cattle and human have similar distributions. For example, both cattle and human phastCons score distribution plots exhibit a prominent peak near 0 and a smaller peak near 1, suggesting our newly created conservation scores for cattle resemble those previously generated for humans (Figure 2.11 A, B, C, D). The majority of the variants were found within 1Mb to the TSS and CpG island (Figure 2.11 E, F, G, H). Additionally, although in a different order, the top five variant consequences in both species include intron variant, intergenic variant, noncoding transcript variant, upstream gene variant, and downstream variant (Figure 2.11 K, L). The different order likely partly reflects the comparatively poor annotation of genes and transcripts in the cattle versus human genome. The annotation pipeline is further applied in subsequent chapters, and in-depth analysis of the annotation results will be included in Chapter 3 and Chapter 4.

For a clearer demonstration of the Enformer predicted scores, the predicted chromatin immunoprecipitation (ChIP) tracks centered on an example human variant were plotted

and presented in Figure 2.12. The example variant is located at position 55436460 on chromosome 17, with the reference allele of C and the alternative allele of A. According to the Enformer predictions, the alternative allele of this variant exhibited a lower ChIP peak in K562 cells compared to the reference allele. Consequently, the allele change was associated with a possible reduced binding of this transcription factor in this cell type as shown in track D of Figure 2.12.



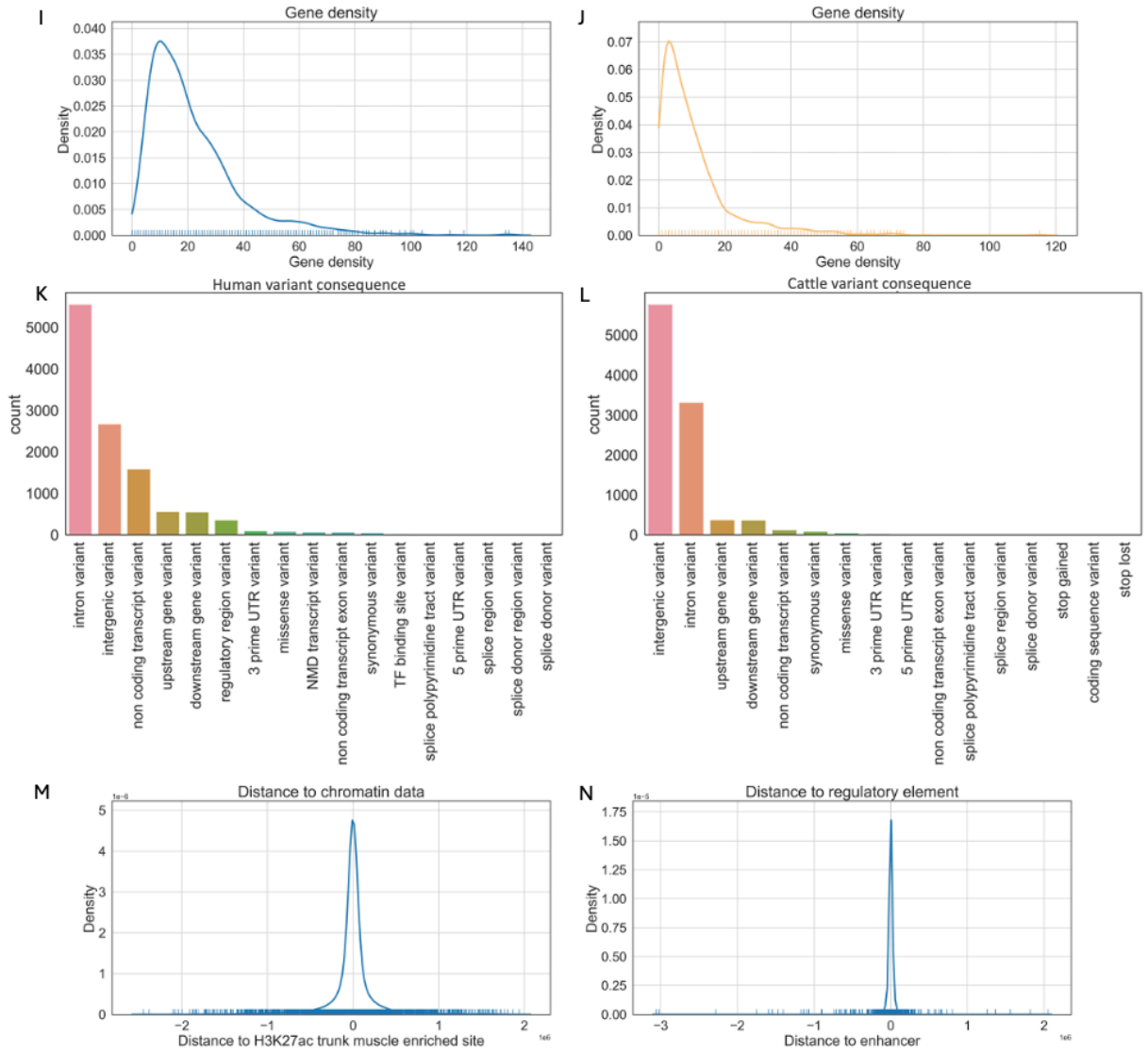



**Figure 2.11 Distribution plots for human and cattle variant annotations.** The human annotation distribution plots are in blue, and the cattle distribution plots are in orange. Each distribution plot is a kernel density estimate (KDE) plot that displays the continuous probability density of the corresponding annotation. The marks along the x axis are the rug plot, where each mark represents a single quantitative variable.

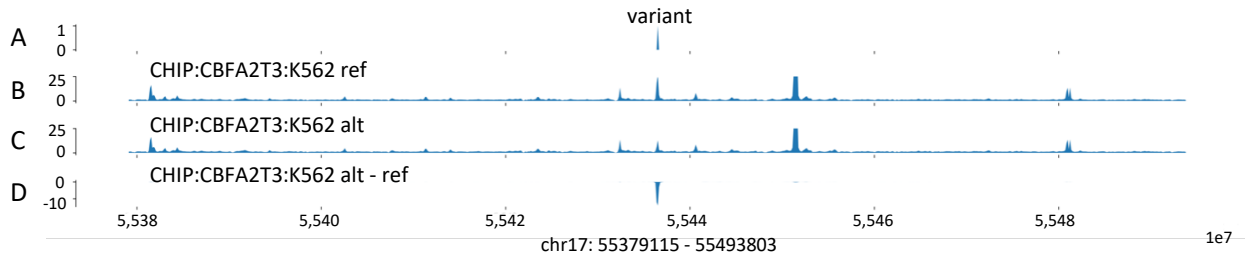



**Figure 2.12 Enformer predicted Chromatin immunoprecipitation (ChIP) track of a human variant chr17:55436460:C:A in K562 cell.** (A) Track shows the position of the target variant. (B) Predicted ChIP track with the reference allele at the variant position. (C) Predicted ChIP track with the alternative allele at the variant position. (D) The predicted difference between the reference and alternative predictions.

## 2.4 Discussion

Variant annotation is an important step in genomic sequence analysis as it provides functional information for variants, which is valuable for downstream variant analysis, including machine learning-based approaches. However, flexible annotation pipelines suitable for machine learning are largely lacking in livestock species. In order to address this gap and facilitate a wide range of downstream projects, a reusable variant annotation pipeline across mammalian species was developed based on Nextflow.

The pipeline incorporated five categories of annotations: sequence conservation, variant position properties, VEP annotation, sequence context and Enformer predicted gene expression score. Sequence conservation annotations included both phastCons and phyloP conservation scores. However, unlike for humans, conservation scores for livestock species are not already available in existing resources. Therefore they were calculated based on the 241-mammlian alignment from the Zoonomia project (Armstrong et al., 2020) and cattle was used as the example species to perform the calculation. Due to the extremely large size of the 241-way multiple alignment file (1.0TB) and the computational complexity involved in the calculation processes, generating conservation scores for livestock species is challenging. Furthermore, the genome assemblies used in the 241-way multiple alignment are not commonly used versions, which means there is a lack of lift-over chain files. This further complicates the generation of conservation scores

and slightly lowers the quality of the data for the current genome assembly. However, it is now feasible to generate the cattle conservation scores for 10,000 variants in two hours using my pipeline. In total, 98.86% of bases in the ARS-UCD1.2 cattle genome are assigned with conservation scores. For other mammalian species, the conservation sub-workflow in the pipeline can be utilized to generate the corresponding conservation scores.

For the variant position properties, distance from the variants to different genome elements, such as CpG island, chromatin data, TSS, and regulatory elements, were included. Due to the lack of source data for livestock species, distance to chromatin data and regulatory elements were only available for human annotations. The popular variant annotation tool VEP was utilized to get the variant consequences and allele frequencies. The 5-mer flanking sequence centered on the target variant was extracted to provide sequence-level information. In addition to these traditional annotations, gene expression scores predicted by the deep learning-based model Enformer were obtained as a set of complementary annotations. The summary of all annotations can be found in Table 2.1. These annotation approaches were then organized into a pipeline using Nextflow as the underlying framework to facilitate the reusability and efficiency of the annotation approaches.

The performance of the variant annotation pipeline was tested on the University server Eddie (*Edinburgh Compute and Data Facility Web Site*, 2021) using different datasets and various computing resources. According to the annotation timelines from various experiments, the pipeline demonstrated the ability to annotate 1000 human variants with basic annotations in 15 minutes and 46 seconds, utilizing 10 cores with 16GB per core. Similarly, the pipeline achieved a comparable elapsed time of 16 minutes and 43 seconds when annotating 10,000 variants using a chunk size of 2,000 using the same computational environment. Additionally, it took 1 hour 44 minutes to obtain the basic features for 10,000 cattle variants using 10 cores with 16GB per core. The Enformer sub-workflow performed much more efficiently when using GPUs compared to CPUs. It took 33 minutes 19 seconds to predict genomic data for 1000 variants using 2 GPU cores with 64GB per core, while it took over 10 hours when using 10 CPU cores with 16GB per core. Furthermore, the conservation score sub-workflow was highly time-consuming, taking

over 24 days to generate the conservation scores using the down-sampled cattle GFF files (40% of the original files) and the MAF files. Note, however, this needs to only ever be done once per species. Consequently, there would likely be merit in using our workflow to generate conservation scores for various livestock species that could then be made available to the wider community.

A limitation of the pipeline is that it provides fewer annotations for livestock species compared to human. This is primarily due to the lack of available source data in appropriate formats, such as chromatin data and regulatory data, in livestock species. With further research on the livestock genome, there is potential for the pipeline to be enhanced and expanded to better annotate livestock variants. For example, although not at the same scale as the human data, the FAANG projects are generating large amounts of omics data for key livestock species including cattle. However, at present they have not been converted into a unified resource of consistently processed features as is available for humans via initiatives such as the Epimap project. Another limitation is that the Enformer sub-workflow currently only supports sequential running. The Enformer model was developed based on Tensorflow, which uses shared global variables to store and update model parameters. As a result, running multiple Enformer models in parallel may lead to conflicts among these shared variables. To address this problem, different approaches have been attempted. One approach is to define separate TensorFlow graphs or sessions for each model, allowing them to run independently without conflicts. However, this introduces compatibility issues between Tensorflow's graph execution and certain functions, such as the .numpy() method in Tensorflow. Another approach is to specify separate GPU or CPU resources for each model, ensuring that they operate in isolation. However, this approach may not be feasible in certain environments due to limited permissions, as in the case of our server Eddie. Further research and development are required to find more effective solutions to achieve the parallel execution of the Enformer models in the Nextflow framework.

Compared to other popular variant annotation tools such as ANNOVAR and VEP, my pipeline offers a wider range of annotations that evaluate the positional properties of the variants by calculating their distances from important genomic elements. Additionally, as

mentioned before, the pipeline has incorporated annotations for other mammalian species, including the novel cattle conservation scores. Furthermore, since the pipeline was built based on Nextflow, it offers greater flexibility in integrating new annotations compared to traditional annotation tools, thereby enhancing its adaptability.

In conclusion, this chapter presented the design and development of a reusable variant annotation pipeline based on Nextflow that can be applied across different mammalian species. The performance of the pipeline was evaluated, highlighting its efficiency and effectiveness. It should be noted that this chapter primarily focused on demonstrating the performance of the pipeline rather than conducting in-depth analysis of the annotation results. Subsequent chapters have utilized the pipeline to generate features for machine learning approaches, delved into more comprehensive analysis of the annotations and explored the insights provided by the annotations in understanding specific types of variants, such as regulatory variants, in both human and cattle.

## 2.5 Supplementary material

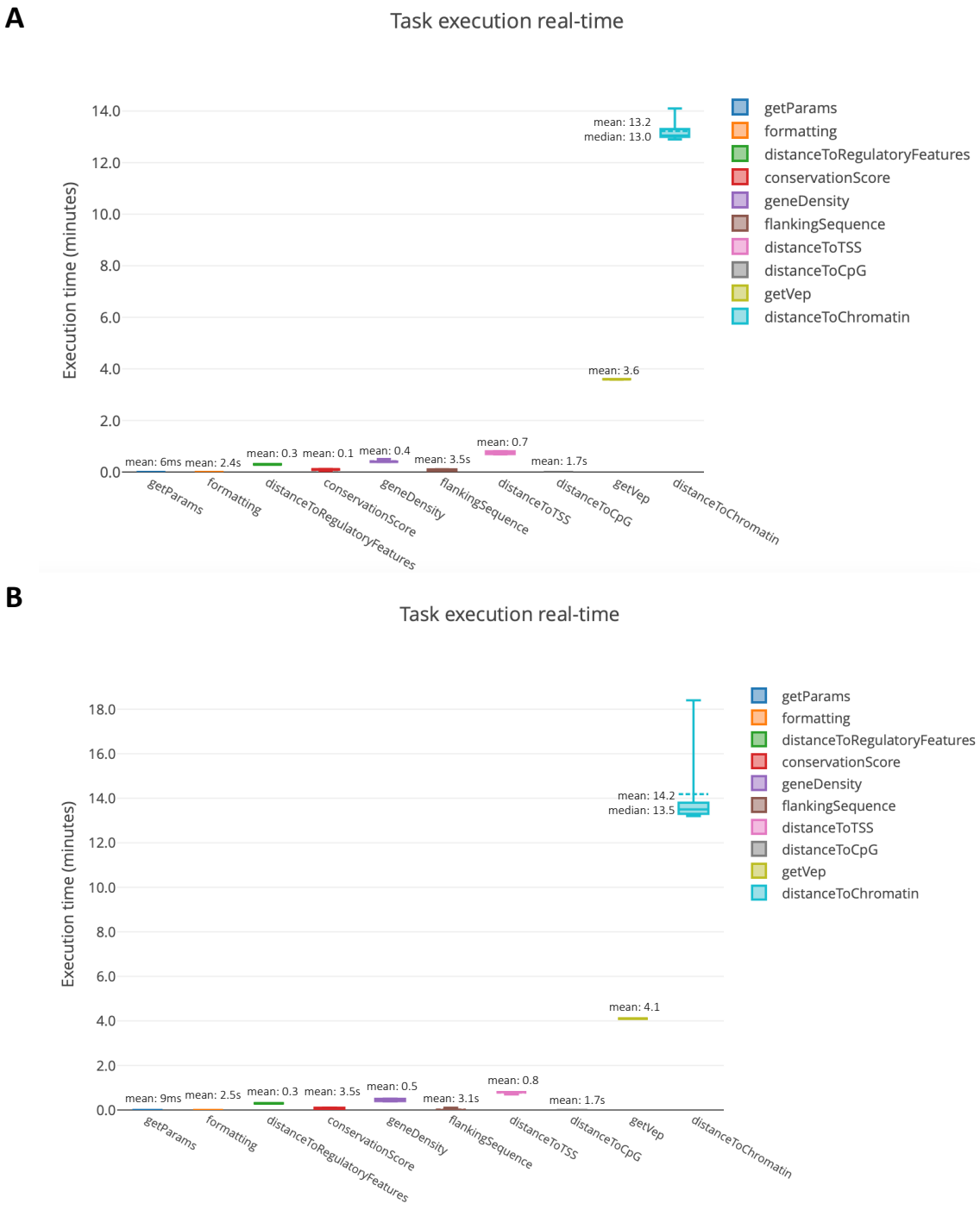



**Figure S2.1 Box plots of processes execution time when annotating 1000 variants using 4 cores with 16GB per core (A) and a single 64GB core (B).**

Elapsed time: 15m 46s
Legend: job wall time / memory usage (RAM)

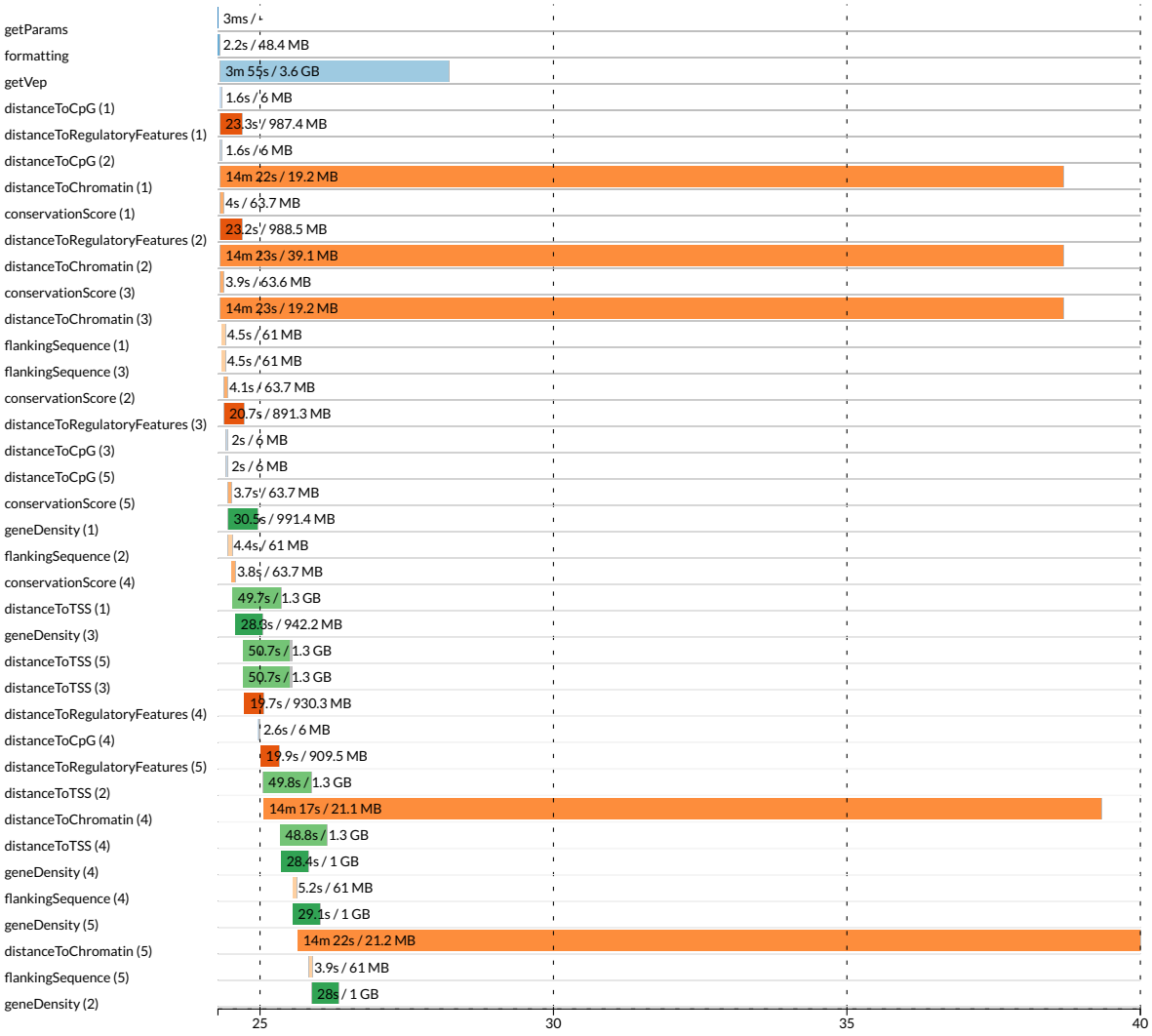



**Figure S2.2 The timeline for all processes executed in the pipeline when annotating 1000 variants using 10 cores with 16 GB per core.**

# Chapter 3 The conservation of human functional variants and their effects across livestock species

## 3.1 Introduction

### 3.1.1 Power of livestock models

Animal models, which refer to living non-human organisms used for studying specific aspects of human biology, play an important role in advancing biological research that is not feasible or ethical to conduct directly in humans. They have proven to be invaluable in different areas, particularly in aiding the understanding of human functional variants that impact downstream phenotypes. In general, small animal models, such as rodents and zebrafish, are currently most often used due to their cost-effectiveness, ease of manipulation and relatively short lifespan. However, these small animal models do have some limitations, including physiological differences compared to humans, which result in the limited translatability of findings to human clinical outcomes (Käser, 2021). Additionally, factors such as evolutionary turnover of regulatory elements and evolutionary distance also make them ill-suited for human biological research (Bourque et al., 2008). An alternative is to use non-human primates, which share a closer evolutionary relationship with humans. However, ethical considerations and high costs limit their use in research (Käser, 2021). Considering these factors, livestock species can potentially serve as better animal models than small animals and non-human primates. Livestock species, such as pig, sheep, and cattle, have already been utilized as animal models in various studies, from toxicology testing of pharmaceuticals to understanding human genetic diseases such as diabetes (Lunney et al., 2021; Pinnapureddy et al., 2015).

### 3.1.2 The issues of genome editing in livestock animal models

Despite the advantages of livestock models, their usage is limited by the practical challenges of conducting genome editing in these species to study the target variants. In

comparison to small animals like rodents, livestock species have much longer generation times. Consequently, it is a very time-consuming process after performing genome editing in livestock species. Additionally, genome-edited animal models introduce artificially engineered variants, which disregard the complex interactions between these variants and other genes and regulatory elements. As a result, these models lack a broader genomic context, potentially limiting the understanding of the comprehensive impact of the human variants.

### 3.1.3 Livestock models with naturally occurring orthologues of human variants

To address the challenges posed by genome editing, an alternative approach is to use livestock models with naturally occurring orthologues of human variants, i.e., variants conserved across species, which is a relatively under-explored strategy. Naturally occurring orthologous human variants refer to mutations occurring and being polymorphic at orthologous positions across species. Livestock models with naturally occurring orthologous variants incorporate the variants into the animal's entire genetic background, facilitating the study of intricate genetic interactions. An example of this is the use of pigs with naturally occurring Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) mutations for studying Cystic Fibrosis (CF) (Meyerholz, 2016). These CF pigs exhibit comparable lung and gastrointestinal symptoms to those observed in human CF patients, making them valuable for gaining insights into the underlying mechanisms of the disease and evaluating potential therapies.

### 3.1.4 Additional advantages of investigating naturally occurring orthologues of human variants

Additionally, investigating livestock with orthologous human functional variants confers significant benefit to the livestock species themselves. Considerably greater biological data and insights have been amassed for humans in comparison to livestock species. The systematic examination of how human functional variants influence corresponding phenotypes in livestock holds the potential to yield valuable findings, contributing to the enhancement of production and health of domesticated animals. For example, the genetic

basis underlying growth and body size overlaps between human and livestock species, such as cattle (Bouwman et al., 2018). The investigation of relevant genetic variants, such as those in the Insulin-like Growth Factor 1 (IGF-1) gene, can contribute to a better understanding of this trait and serve as guidance for downstream breeding (Bouwman et al., 2018; Mullen et al., 2011). Another example is the similarities in the genetic basis of immune response and disease resistance between humans and pigs (Pabst, 2020). The investigation of genetic variants associated with immune response can help improve the ability to combat infectious diseases in livestock species, subsequently leading to an improvement in their welfare. Overall, investigating the conservation of human functional variants across livestock species aligns with the principles of One Health, which seeks to recognize the interconnection of human and animal health.

Another potential advantage of studying the naturally occurring orthologues of human variants across species lies in its ability to offer insights into the feasibility of extending the utilization of limited high-quality data across diverse species in advanced computational approaches such as machine learning. Machine learning approaches have proved to be useful for predicting functional variants and their effects in humans as discussed in chapter 1.2.3.4. However, the adoption of machine learning in livestock species is rare, primarily due to the lack of reliable training datasets. By exploring the conservation of functional variants across species, it is possible to leverage the larger and higher-quality human data to other species. Additionally, it enables the application of transfer learning in developing machine learning models in livestock species, which allows knowledge gained from one domain, i.e., humans, to be transferred and utilized in another domain, i.e., livestock species.

### 3.1.5 Genomic datasets

To date, there has been limited exploration of natural orthologues of human functional variants and the extent to which their effects are conserved across mammalian species. This is partly attributed to the lack of high-quality datasets containing precise livestock functional variants and the absence of reliable livestock genomic data. With the

improvement of the datasets, the expansion of the exploration to genome-wide level is both feasible and necessary.

The 1000 Genomes Consortium provides a comprehensive collection of common human genetic variations, which encompasses over 78 million human SNPs, across 2504 individuals from 26 different populations (Auton et al., 2015). Similar genotype datasets, though on a smaller scale, have been generated for livestock species. For example, Dutta et al. identified 26,247,559 biallelic single nucleotide variants (SNV) across 79 water buffalos and 477 cattle, (Dutta et al., 2020). Li et al. developed the Genome Variation Map, which incorporates genotypes for 41 different species including pig, cat (C. Li et al., 2021). These genotype datasets make it possible to investigate common variants across different species. Furthermore, various human functional variant datasets, such as the pathogenic variants from the Clinvar database (Landrum et al., 2018), the common variants linked to polygenic diseases and traits from the UK Biobank cohort (Weissbrod et al., 2020), and the fine-mapped regulatory variants from the GTEx project, enable the exploration into the conservation of these functional variants across species. Additionally, with the extension of the GTEx projects from humans to livestock species such as cattle GTEx (S. Liu et al., 2022), further comparison of the effects of the regulatory variants across species, both in terms of the association between the variants and the target genes, and the direction of the effect, becomes feasible.

### 3.1.6 Livestock orthologous variants annotation and prediction

Following genome-wide investigation of the conserved variants across species, it is also necessary to explore the underlying characteristics associated with those shared variants. The variant annotation pipeline developed in Chapter 2 can be utilized to annotate the human variants with or without orthologues in other livestock species. These annotations can provide insights into potential factors associated with conserved variants, such as their distances to different genomic elements and the sequence information of the variants. Furthermore, these annotations can serve as features in machine learning approaches, facilitating investigations into the feasibility of predicting the presence of a livestock orthologue for a given human variant. In addition to the prediction aspect, certain

machine learning approaches can offer valuable insights into feature importance, thereby enhancing understanding of the predominant factors influencing the prediction of conserved variants.

### 3.1.7 Objectives

The objective of this chapter was to assess the prevalence of natural orthologues of human variants in domesticated species. The genotype datasets for human and four different domesticated species including cattle, pigs, dogs, and water buffalo were first obtained to perform genome-wide exploration of shared variants across species. Employing machine learning approaches, discussed in Chapter 1, we sought to characterize the features related to the presence of orthologues across different species. Furthermore, the presence of functional variants across mammalian species, including pathogenic variants, common variants associated with polygenic disease and traits, and regulatory variants, were investigated and the conservation of their effects on downstream phenotypes assessed. We emphasized the potential significance of human functional variant orthologues as a valuable resource for enhancing the understanding of the genetic basis underlying both human and livestock phenotypes.

## 3.2 The conservation of human functional variants and their effects across livestock species

Rongrong Zhao[1], Andrea Talenti[1], Lingzhao Fang[1], Shuli Liu[2], George Liu[3], Neil P Chue Hong[4], Albert Tenesa[1], Musa Hassan[1], James G.D. Prendergast[1]*

[1] The Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, United Kingdom
[2] Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang 310024, China
[3] Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service, Agricultural Research Service, USDA, Beltsville, Maryland 20705, USA
[4] EPCC, Bayes Centre, 47 Potterrow, Edinburgh, EH8 9BT, United Kingdom

*To whom correspondence should be addressed: (james.prendergast@roslin.ed.ac.uk).

**Abstract**

Despite the clear potential of livestock models of human functional variants to provide important insights into the biological mechanisms driving human diseases and traits, their use to date has been limited. Generating such models via genome editing is costly and time consuming, and it is unclear which variants will have conserved effects across species. In this study we address these issues by studying naturally occurring livestock models of human functional variants. We show that orthologues of over 1.6 million human variants are already segregating in domesticated mammalian species, including several hundred previously directly linked to human traits and diseases. Models of variants linked to particular phenotypes, including metabolomic disorders and height, are preferentially shared across species, meaning studying the genetic basis of these phenotypes is particularly tractable in livestock. Using machine learning we demonstrate it is possible to identify human variants that are more likely to have an existing livestock orthologue, and,

importantly, we show that the effects of functional variants are often conserved in livestock, acting on orthologous genes with the same direction of effect. Consequently, this work demonstrates the substantial potential of naturally occurring livestock carriers of orthologues of human functional variants to disentangle their functional impacts.

**Introduction**

Animal models are widely used across the biological sciences. From the development of vaccines and use as models of human diseases, to addressing fundamental questions about human biology. Importantly animal models provide the ability to test the effect of manipulating key variables in a controlled fashion, in ways that are not possible in human populations. For example, by altering the genome via genome editing. The introduction of variants thought to be functional in humans into animal models enables a range of studies, from the characterization of their downstream impacts on the expression of genes, to how different alleles respond to different interventions such as drug treatments.

By far the most widely used mammalian animal models are rodents, due to their ease of handling and short generation times. But rodent models have several limitations. Most importantly humans and rodents are physiologically very different, with the pathogenesis of diseases often differing substantially between the species. This has been proposed as a key driver of why less than 8% of cancer studies that are based on animal models result in a clinical trial[1]. Furthermore, the sizes of rodent organs poorly match those of humans, and it is difficult to serially sample rodent models due to their smaller size. Although the use of primate models can overcome many of these limitations their use is limited by both cost and ethical considerations[1]. For these reasons livestock species have been proposed as more effective animal models in many scenarios[2,3]. Pigs in particular have a similar size, physiology and anatomy to humans[4], and have been shown to have more similar gene expression patterns to humans than rodents[5]. As a result they are increasingly used in translational research, from toxicology testing of pharmaceuticals to the development of transgenic models of human diseases ranging from cystic fibrosis and diabetes to neurodegenerative disorders[6]. However, livestock models of human functional genetic variants have major drawbacks: they are expensive and time-consuming to

generate. As well as the substantial time and costs associated with generating and implanting the genome edited embryos, it is necessary to maintain the mothers through long pregnancies in areas suitable for genetically modified animals, with no prospects of recouping the costs through selling the animals afterwards. There are also further ethical considerations to such transgenic projects, with the public often skeptical of the merits of artificially introducing human variants into other species.

Therefore, despite the clear merits of being able to assay the effects of human functional variants in livestock models, transgenic experiments come with several obstacles. Even among mice, the number of truly humanized models, i.e., where the directly orthologous mouse base or sequence has been altered to match that in humans, is low. Traditionally transgenic mouse models involve the random insertion of transgenes into the genome, meaning they lose their wider genomic context and potential impacts on downstream functions and mechanisms[7]. To properly model human functional variants, the same changes need to be made at orthologous locations, with both alleles present among the animal model.

A relatively under-explored alternative to the *de novo* generation of animal models is the study of natural orthologues of human functional variants. The 1000 Bull Genomes Project alone identified over 84 million cattle single nucleotide polymorphisms[8], meaning approximately 1 in every 32 bases in the cattle genome is polymorphic. This though is potentially an underestimate of the expected probability of a human variant having a cattle orthologue, as polymorphisms are known to be dependent on the underlying sequence. For example, CpG sites are known to be susceptible to deamination, likely raising the probability of such sites being polymorphic across species. This suggests there are potentially many natural orthologues of human functional variants, meaning the effect of these variants can be studied in large mammalian models, potentially at scale, without resorting to transgenic approaches. Supporting this idea, although rare in the literature, some examples of functional variants being found naturally across different mammalian species have already been reported. For example, a missense change linked to coat colour found segregating among both dogs and water buffalo[9]. In recent work, non-

naturally occurring coding changes in mice and zebrafish were compared to these found in humans, with orthologues of human pathogenic Clinvar variants shown to more likely also to lead to a detectable phenotypic change in zebrafish than other variants[10]. To date there has though been little genome-wide study of the natural orthologues of human functional variants and the conservation of their effects across mammals. In part this has resulted from the fact that the precise functional variant underlying most human quantitative trait loci and genome-wide association loci have been unknown. However, high resolution functional datasets and fine-mapping approaches have begun to disentangle causative variants from those simply in linkage disequilibrium[11,12].

Studying the impact of these functional variants has the potential to inform our understanding of phenotypes beyond just humans. This is because livestock species are not only good models for humans but the reverse is also true. Substantially more biological data and insights have been generated for humans than livestock, and characterizing how human functional variants effect corresponding phenotypes in livestock may provide insights into how to improve the production and health of domesticated animals. For example, the genetic basis of stature in cattle has already been shown to have parallels of that in humans[13,14], and better understanding functional variants linked to height could provide potential avenues for adjusting livestock body size.

The aim of this study was, therefore, to characterize the extent to which natural orthologues of human variants are found in domesticated species. Using machine learning we characterize the features associated with the presence of orthologues across species, investigate the presence of functional variants linked to diseases and traits across mammals, and determine where their effects on downstream phenotypes are conserved. We highlight how orthologues of human functional variants are likely a valuable resource to better understand the genetic basis of both human and livestock phenotypes.

## Results

### *Extensive sharing of variants across species*

To investigate how often the same variants are found across species, we compared the 78 million human SNPs identified in the 1000 genomes cohort of ~2500 diverse individuals[15] to the variants identified in cohorts of 477 cattle[9] and 409 pigs[16]. In total 35 and 34 million of the human variants could be mapped to an orthologous location in the pig and cattle genome, respectively (Figure 1a). Of these 3.7 and 3.0% overlapped an orthologous variant segregating in one of these other species, with 55.5 and 55.8% of these showing the exact same allele change. Consequently over 1.1 million human variants have a direct orthologue in at least one of these two livestock cohorts. Intersecting the same human polymorphisms with variants in cohorts of two further domesticated species, 722 dogs[17] and 81 water buffalo[9], revealed that 1,651,728 are found in at least one of these four mammalian cohorts (Figure 1b).

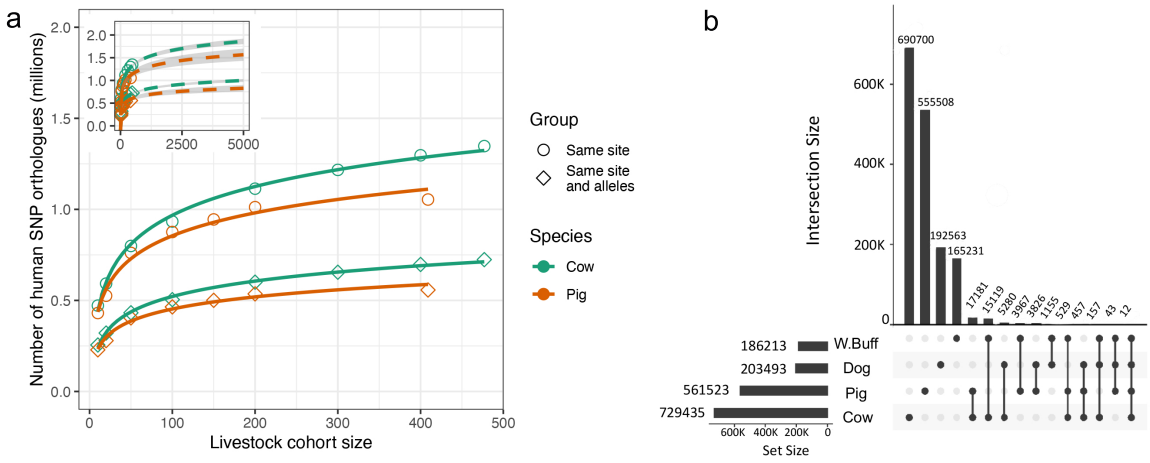



**Figure 1. Frequency of variant sharing across species.** (a) Number of human (1000 genomes) SNPs that have a SNP at the orthologous location in each other species. Counts are broken down into where the SNPs have the same alleles across species (same site and alleles) or simply coincide, i.e. irrespective of allele change. The inset shows the number of orthologous SNPs expected in larger cohorts when extrapolating the curves. (b) The number of human variants overlapping a variant found in one or more other species with a matching allele change. The vertical bars indicate the number of human orthologues found across the species indicated by the dots below. The horizontal bars indicate the total number of orthologues in that species. The data underlying this plot can be found in Supplementary Data 1.

The number of variants shared across cohorts from different species is expected to be a function of the number of samples in each cohort. To characterize this relationship we randomly down-sampled the pig and cattle cohorts and recalculated the observed overlap with the total set of human variants. As shown in Figure 1a the number of variants overlapping the human dataset had not plateaued for either species, suggesting larger cohorts would continue to identify even more orthologues of human variants. For example, extrapolating the results to 5000 samples in corresponding cohorts suggests over 840,000 pig and 1,000,000 cattle orthologues of human variants would potentially be detected (Figure 1a). As expected, sample diversity/relatedness is also an important factor with more diverse cohorts leading to more orthologous variants being identified (Supplementary Figure 1). This suggests exact orthologues of several million human variants are naturally segregating among livestock species.

***Modelling the distribution of shared variants across the genome***

Using 1589 different human annotations (Table 1), including sequence conservation, chromatin context, and the distance to genome features such as genes, we investigated whether these features were linked to whether human variants had livestock orthologues. Several factors were associated with the probability of a human SNP having a cattle orthologue, including their distance to known genes and chromatin, and sequence context. For example, human variants with a cattle orthologue are more likely to involve a C to T change than those without a corresponding cattle orthologue (Figure 2a). C to T changes in mammalian genomes are commonly caused by the known hypermutability of CpG sites, whereby CpG sites are highly susceptible to deaminate to TpG[18]. The elevated mutation rates of these sites consequently likely increases the chance of the same change occurring across lineages. More generally, human changes with a G:C base pair within their 5-mer flanking sequence are more likely to have a cattle orthologue than those with an A:T base pair at the same position (Figure 2b). A notable exception to this is where a guanine is found 5 prime of the human SNP site, with such changes less likely to have an orthologous SNP at the same position in cattle (Figure 2b). Variants with orthologues

are also more likely to be enriched near specific genes such as processed pseudogenes and snoRNA, and around certain chromatin marks (Figures 2c, d).

| | Annotation | Data source | Encoding method | Number of features | Number of columns after encoding |
|---|---|---|---|---|---|
| Sequence conservation | phastCons100way | UCSC genome annotation database | - | 1 | 1 |
| | phastCons30way | | - | 1 | 1 |
| | phyloP100way | | - | 1 | 1 |
| | phyloP30way | | - | 1 | 1 |
| Variant position properties | Distance to CpG island | UCSC genome annotation database | - | 1 | 1 |
| | Distance to chromatin data [a] | Ensembl | - | 1554 | 1554 |
| | Distance to TSS [b] | Ensembl | - | 14 | 14 |
| | Distance to regulatory features [c] | Ensembl | - | 4 | 4 |
| | Chromosome | - | One-hot encoding | 1 | 22 |
| | Variant position | - | - | 1 | 1 |
| | Gene density (per megabase) | Ensembl | - | 1 | 1 |
| VEP annotations | Consequence | Ensembl | One-hot encoding | 1 | 34 |
| | Allele frequency [d] | | - | 6 | 6 |
| Sequence context | Allele change | Ensembl | Self-defined encoding | 1 | 8 |
| | 5-mer flanking sequence | UCSC genome annotation database | | 1 | 20 |

[a] *Distance to 1554 different chromatin data types from Ensembl (regulatory build of hg38).*
[b] *Distance to TSSs within 14 common biotypes (e.g. protein coding, lncRNA etc. Those with a frequency in the genome of >= 1000).*
[c] *Regulatory features include enhancer, promoter, CTCF binding site, TF binding site.*
[d] *A total of six allele frequencies from the 1000 genomes combined population and the African, American, East Asian, European and South Asian populations separately.*

**Table 1. Variant annotations used in the modelling and their encoding method.** In total 1589 features were used in the machine learning models to assess whether they were linked to the probability of a variant being found across species. These broadly fell into one of four categories and after encoding (conversion of categorical data to integer format) there were a total of 1669 features tested in the modelling.

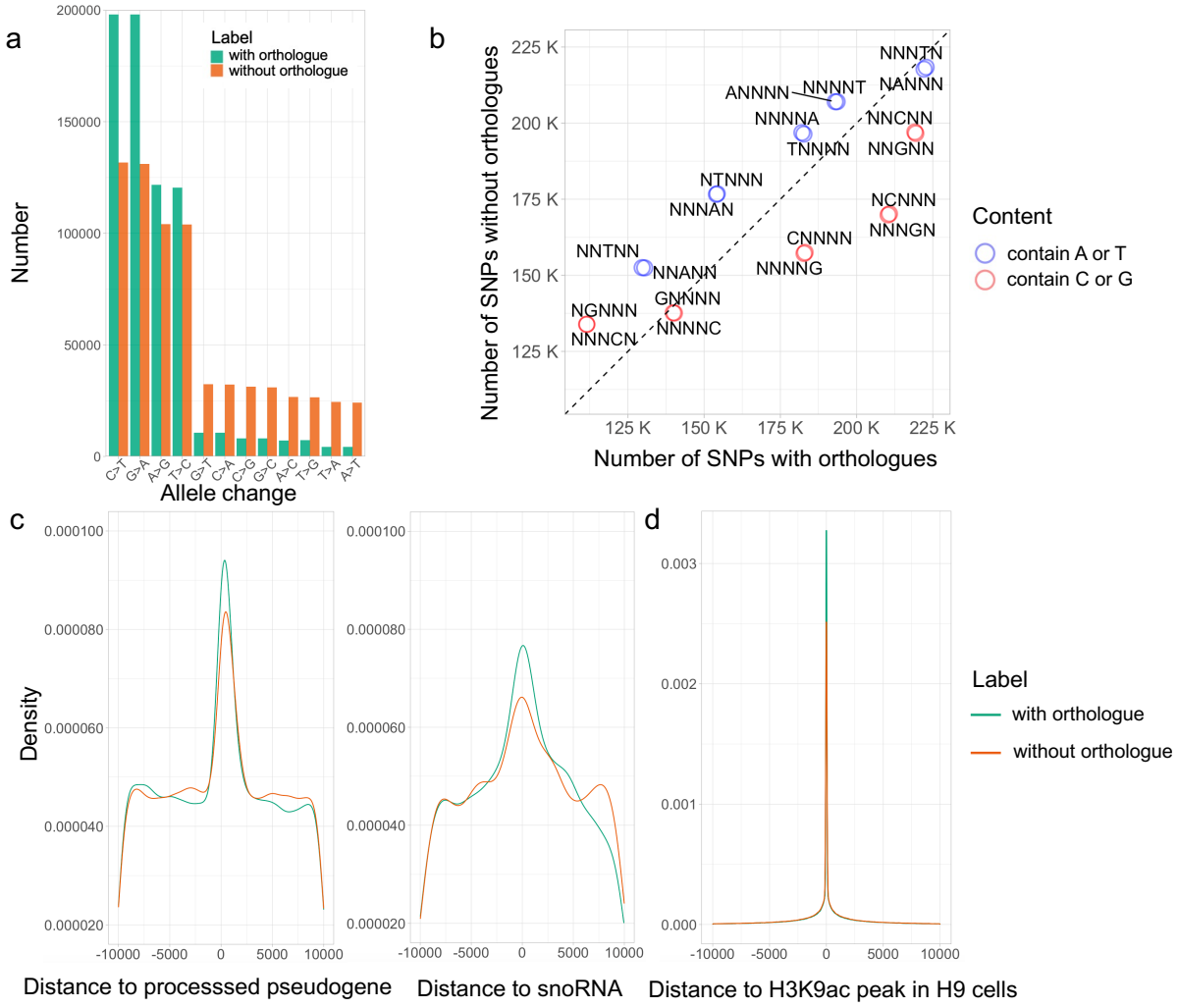
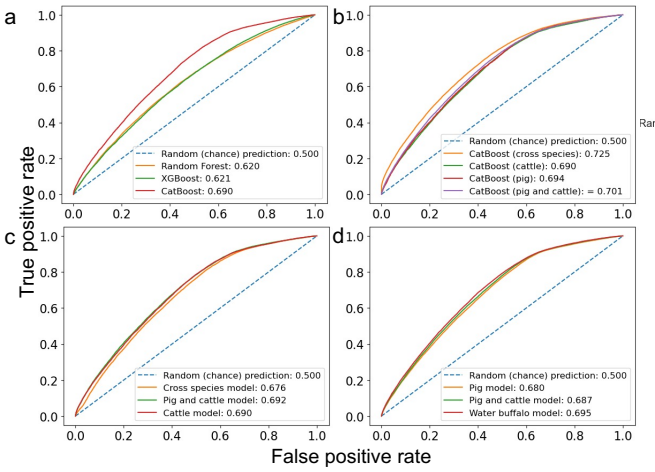
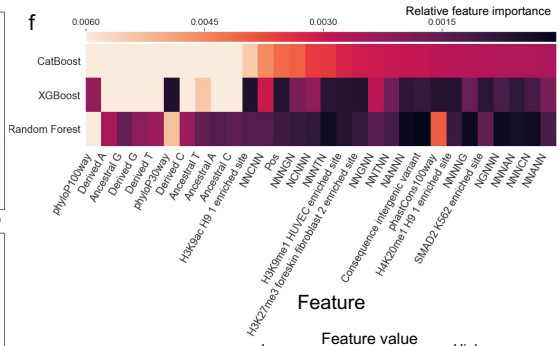
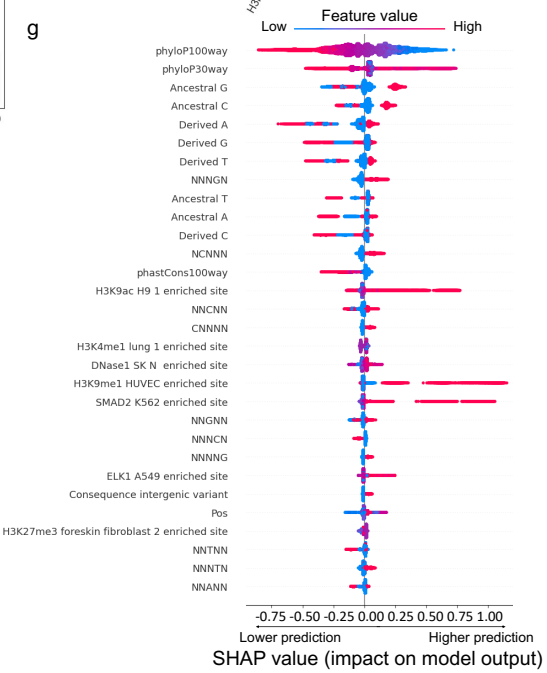



**Figure 2. The characteristics of human variants with cattle orthologues.** The association of 1,395,750 human variants with different genomic features is shown (697,875 human variants with orthologues in cattle and an equally sized random sample of 697,875 human variants without known orthologues in cattle). (a) Number of variants with or without orthologues by their observed allele changes (reference > alternative). The data underlying this plot can be found in Supplementary Data 2. (b) Number of SNPs with different 5-mer flanking sequences among variants with or without orthologues. Each circle represents a 5-mer flanking sequence with a specific base at a certain position, and the circle colour indicates whether the specific base is C/G or A/T. The black dashed line represents parity, i.e., where the number of SNPs with orthologues equals the number of SNPs without known orthologues. All the 5-mer sequences are significantly different between the groups at a p-value less than $2.2 \times 10^{-16}$ (Chi-Squared test). (c) Density plots of distances of variants with or without orthologues to processed pseudogenes and snoRNAs (plot restricted to within 10kb). Distances of variants to processed pseudogenes and snoRNAs are different between groups at p-values less than $3.2 \times 10^{-5}$ (Two-sample Kolmogorov-Smirnov test). (d) Density plot of distance between variants with or without orthologues to chromatin regions marked by H3K9ac in the human H9 cell line (plot restricted to within 10kb). Distance to these regions is different between groups at p-value less than $1.8 \times 10^{-3}$ (Two-sample Kolmogorov-Smirnov test). The data underlying this plot can be downloaded from https://doi.org/10.6084/m9.figshare.20401851.

We investigated the extent to which it is possible to use these genomic annotations together to predict whether a human variant will have an orthologue in a livestock species. To do this, we used 140,000 human variants with or without a cattle orthologue and trained three tree-based machine learning models (Random Forest, XGBoost and CatBoost, see methods) on the 1589-human genomic features (Table 1). To compare the performance of these models at discriminating human variants with and without cattle orthologues we tested the models on a further 60,000 human variants that had not been included in the original training data. As shown in Figures 3a and 3e, CatBoost outperformed the other models with an area under the receiver operating characteristics (AUC) score of 0.69, an accuracy of 0.64 and an F1 score of 0.70. These results suggest that the genomic annotations of the variants contain discriminating information that makes it possible to identify which human variants have a higher probability of having an orthologue in another species.




e

| Model | Training Data | Test Data | Accuracy | F1 score | AUC |
|---|---|---|---|---|---|
| Random Forest | | | 0.60 | 0.62 | 0.62 |
| XGBoost | Cattle | Cattle | 0.61 | 0.65 | 0.62 |
| CatBoost | | | 0.64 | 0.70 | 0.69 |

| Model | Training Data | Test Data | Accuracy | F1 score | AUC |
|---|---|---|---|---|---|
| CatBoost | Pig, cattle, dog, water buffalo | Pig, cattle, dog, water buffalo | 0.67 | 0.71 | 0.73 |
| | Cattle | Cattle | 0.64 | 0.70 | 0.69 |
| | Pig | Pig | 0.64 | 0.70 | 0.69 |
| | Pig, cattle | Pig, cattle | 0.65 | 0.71 | 0.70 |

| Model | Training Data | Test Data | Accuracy | F1 score | AUC |
|---|---|---|---|---|---|
| CatBoost | Pig, cattle, dog, water buffalo | Cattle | 0.62 | 0.68 | 0.68 |
| | Pig, cattle | | 0.64 | 0.69 | 0.69 |
| | Cattle | | 0.64 | 0.70 | 0.69 |

| Model | Training Data | Test Data | Accuracy | F1 score | AUC |
|---|---|---|---|---|---|
| CatBoost | Pig | Water buffalo | 0.63 | 0.68 | 0.68 |
| | Pig, cattle | | 0.64 | 0.69 | 0.69 |
| | Water buffalo | | 0.65 | 0.70 | 0.70 |

**Figure 3. Machine learning models of orthologous variants.** (a) Receiver operating characteristic (ROC) curves of Random Forest, XGBoost and CatBoost models trained and tested using human variants with and without orthologues in cattle. The numbers in the legend are area under the receiver operating characteristics (AUC) scores of the different models. AUC reflects a model's general ability of distinguishing between the classes, the closer to 1 the better the model performance. Values of 0.5 would reflect a model not able to differentiate between variants with and without orthologues. (b) ROC curves of CatBoost models trained and tested using human variants with and without orthologues in cattle; pig; pig or cattle; pig, cattle, dog or water buffalo (cross species). (c) ROC curves of CatBoost models trained using human variants with and without orthologues in cattle; pig or cattle; pig, cattle, dog or water buffalo, but tested using human variants with and without orthologues in cattle. (d) ROC curves of CatBoost models trained using human variants with and without orthologues in pig; water buffalo; pig or cattle, and tested using human variants with and without orthologues in water buffalo. (e) Summary statistics of the experiments. Tables correspond to the same analyses shown in plots A to D respectively in the same order. (f) Feature heatmap of CatBoost, XGBoost, Random Forest models trained and tested using human variants with and without orthologues in cattle. Thirty important features are included in the figure, with lighter colour indicating greater importance in that model. (g) SHAP summary plot[19] of the CatBoost model trained using human variants with and without orthologues found in any of cattle, pig, dog and water buffalo. A SHAP plot shows the relationship between feature values and their impact on model predictions. Features are ranked in descending order according to their importance on the left. The colour represents low (in blue) and high (in red) value of the feature and the effect of their values on the output of the model is reflected by their positions on the x-axis. For example, lower phyloP100Way conservation scores (as indicated in blue) are associated with an increased probability of a human variant having an orthologue in another species (being higher on the x axis). The SHAP values underlying this plot can be downloaded from https://doi.org/10.6084/m9.figshare.20730370. The data underlying Figure 3 can be downloaded from https://doi.org/10.6084/m9.figshare.20401851.

Models trained on human variants with orthologues in other species, such as pig, could predict the presence of orthologues in these other species with similar accuracies as the cattle specific models (Figures 3b,c,e). Likewise models trained on human variants with orthologues in given species were largely as accurate at predicting orthologues in completely different species (Figures 3d,e). This suggests the features associated with orthologous variants are fundamental across mammals.

Comparison of the top 30 most important features of the three different modelling approaches shown in Figure 3a found that the allele change, 5-mer flanking sequence and conservation score (phyloP100way) were consistently three important features (Figure 3f). Figure 3g shows the top 30 most important features of the cross-species model, i.e. trained using human variants with an orthologue in any of the tested livestock species, and how their values affect the predictions of the model. Sequence conservation is the most important variable, with human variants in less-conserved regions more likely

to have an orthologue in another species. This is consistent with mutations in these regions less often being removed, increasing the probability of the same change occurring in different mammalian lineages. As well as the type of base change the flanking sequence disproportionately contributes to the model performance, with a G base at the position immediately downstream of the polymorphic site (NNNGN) associated with an increased probability of the same change being observed in the other species, consistent with the preferential deamination of CpG sites.

***Animal models of human pathogenic variants***

Although over a million human variants have a livestock orthologue (Figure 1), the modelling results above highlight that these disproportionately fall in less conserved, and consequently most often non-functional, genomic regions. This raises the question as to how many naturally occurring livestock models of functional human pathogenic variants exist. To characterize specifically how many human pathogenic mutations are segregating in other livestock species we first extracted 89,158 SNPs from the human Clinvar[20] database labelled as “pathogenic” or “likely pathogenic”. Being mostly found in conserved coding regions, the overwhelming majority (99.4% and 94.1%) of these variants could be successfully mapped to an orthologous position in the Cow (BosTau9) and Pig (SusScr3) genomes. Using the data from the same cow and pig cohorts we identified how often these variants overlapped an orthologous variant segregating in one of these other species. In total 1290 Clinvar variants overlapped a variant in the cow dataset and 767 in the pig cohort, of which 253 and 212 respectively also showed the exact same allele change observed in humans. In agreement with the modelling results, these numbers differ from those expected from the background numbers. Not only is the number of Clinvar variants with an orthologue in one of these livestock species substantially lower than expected given the number of all human variants with an orthologue in pig or cattle, but also where a variant does segregate at the orthologous position it is less likely to show the same allele change. In total Clinvar variants are approximately three times less likely to have a variant at the orthologous position in either pig or cattle than expected from the frequencies in the 1000 genomes cohort, and approximately seven times less likely of

having one displaying the same allele change (Supplementary Figure 2). These results are consistent with these changes being deleterious as indicated, and selection preferentially removing them across species.

Orthologous variants, even with the same allele change, may not have the same impact on genes, if for example the gene structure and codons have changed between species. As shown in Figure 4a, 80% of the 103 cattle orthologues of human Clinvar variants leading to a missense change show the same missense change across both species. A further 13% are missense in both species but involving different amino acid changes. Only 3.9% of the human missense variants are predicted to be synonymous in cattle, suggesting the consequence of human missense changes is most often conserved across these mammals, with similar patterns observed in pigs (Figure 4b). However, of 115 human Clinvar variants predicted to lead to the introduction of a stop codon, only 22% also lead to a stop gained change in cattle, with the majority (63%) predicted to just lead to an amino acid change due to a difference in the codon between species. This may represent a true difference in the impact of these variants between species but may also sometimes reflect the comparatively poor annotation of gene isoforms in livestock species. Of note, 33 cattle and 20 pig variants lead to the same protein impact as their orthologue in humans despite involving a different allele change (Figures 4c,d). Consequently, although rare, variants do not necessarily need to show the same allele change to have a conserved impact.

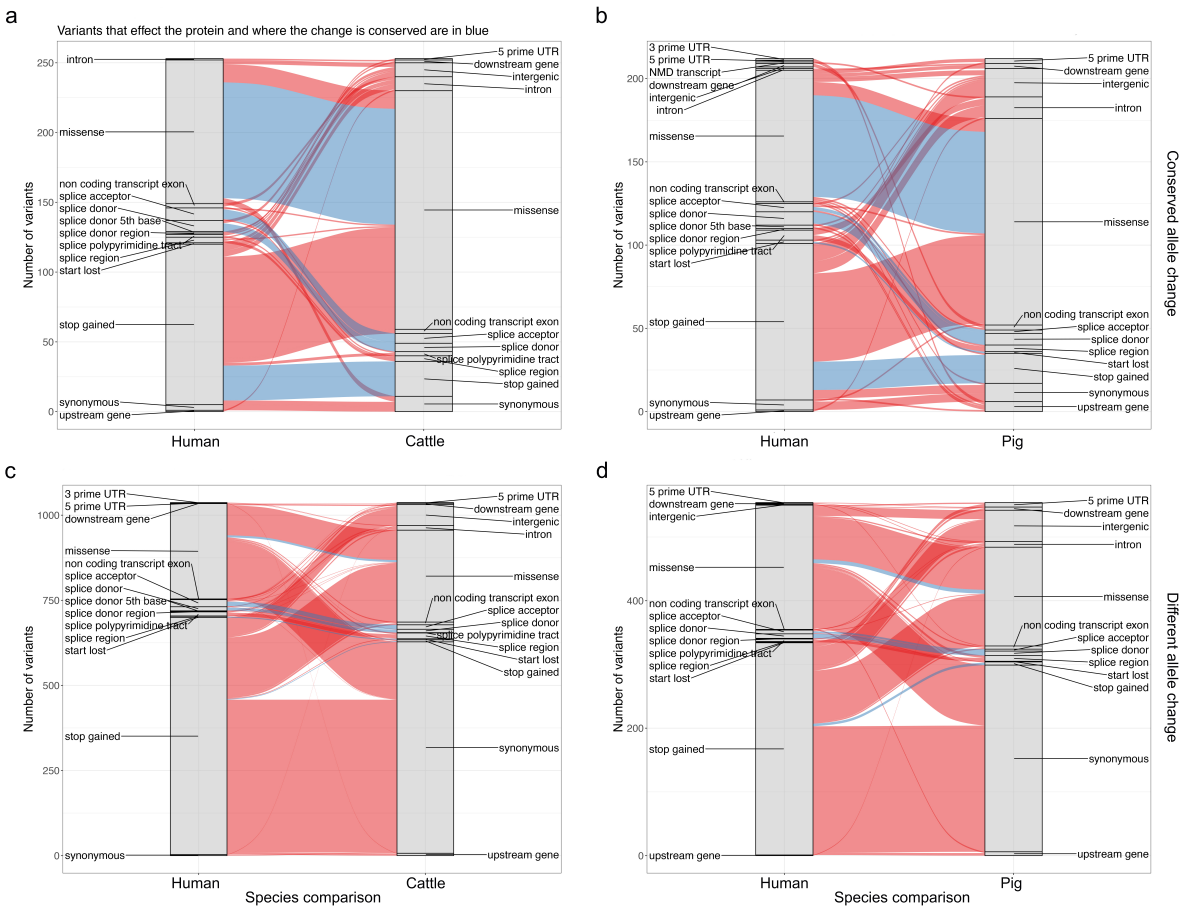



**Figure 4. Conservation of impacts of orthologues of Clinvar variants**. Each plot shows the consequence of variants in humans (on the left) and the consequence of variants found at the same position in the pig or cattle genome on the right. The ribbons connect sets of variants, with the width of each ribbon indicating the number of variant pairs with the given combination of consequences across the two species. Ribbons in blue indicate where variants have potentially conserved impacts on the protein across species. (a) The conservation of impact of genic orthologous variants across human and cattle where variants show the same allele change. (b) The same as (a) but for human-pig orthologous variants. (c) Conservation of impact of variants across human and cattle where their locations are orthologous but they show different allele changes. (d) The same as (c) but for human-pig orthologous variants. The data underlying this figure can be found in Supplementary Data 3.

These data highlight that there are existing animal models available for at least several hundred human Clinvar variants, including those linked to a variety of important phenotypes such as cancers and Parkinson's disease (Supplementary Data 4). Interestingly, Clinvar variants linked to certain traits are more likely to be found across species. This includes those linked to biotinidase deficiency (Chi-squared test $P < 1x10^{-7}$), neurofibromatosis ($P=1.8x10^{-5}$) and glycogen storage in cattle ($P=3x10^{-7}$) and factor

VII deficiency in pigs ($P < 1x10^{-7}$). Of 23 known human biotinidase deficiency variants, four (17%) have a direct orthologue showing the same allele change in cattle. This is despite only 1.5% of all lifted Clinvar variants having a cattle orthologue. All four of these variants are missense SNPs showing the same amino acid change in both species, with one of the mutations having risen to a minor allele frequency of 22% in cattle despite being found at a frequency of only 0.002% in humans, meaning studying its impact may be easier in cattle populations. Supplemental biotin is often fed to cattle as it is thought to improve hoof health and increase milk production[21] and these variants are consequently also strong candidate functional variants for further investigations for improvement of these cattle traits.

***Animal models of common variants of polygenic diseases and traits***

We next investigated whether there are potential existing livestock models of common human variants linked to polygenic diseases and traits. To do this, we obtained 2240 fine-mapped SNPs linked to 47 different traits in the UK biobank cohort[22]. In total 58 of these variants had a direct orthologue segregating in either pigs or cattle. Interestingly variants linked to height in humans were significantly more likely to have a direct orthologue in cattle than other traits, with over a quarter of variants (11 out of 43) that were found in both species with the same alleles being linked to this phenotype (Figure 5, Supplementary Data 5). This is compared to only 13.6% (341 out of 2513) of the variants successfully mapped between the species being linked to this trait (two-tailed Fisher's exact test P=0.040). Of these 11 variants, 3 are missense changes (rs154001, rs61735104, rs79485039), with each leading to the same amino acid change in both species. These amino acid changes fall in *FGFR3*, *KIAA1614* and *FBN2*, with a further gene, *FOXM1*, having a variant (rs28990715) at orthologous positions in both species that leads to the same amino acid change despite having different allele changes. 10 of the 11 human variants with cattle orthologues were in the Gene Atlas UK Biobank[23] results and are together, under certain assumptions, associated with a predicted 2.7cm variation in human height. This corresponds to around ~1/3 of a standard deviation of the human heights in the UK Biobank cohort. The FGFR3 change alone is associated with a ~1cm

difference in standing height between opposing homozygotes. Mutations in FGFR3 underlie 99% of cases of human achondroplasia that affects bone development and leads to short stature, but the role of this gene in cattle stature is less well characterized[24]. Potentially, in part, because all 11 variants are rare in cattle (10 with a frequency of less than 1%, with 1 with a frequency of 6%) and would be unlikely to be detected in a standard cattle GWAS, but these variants are consequently strong candidate functional rare variants for contributing to variability in cattle height due to their strong associations with this trait in humans that could be exploited to alter cattle stature.

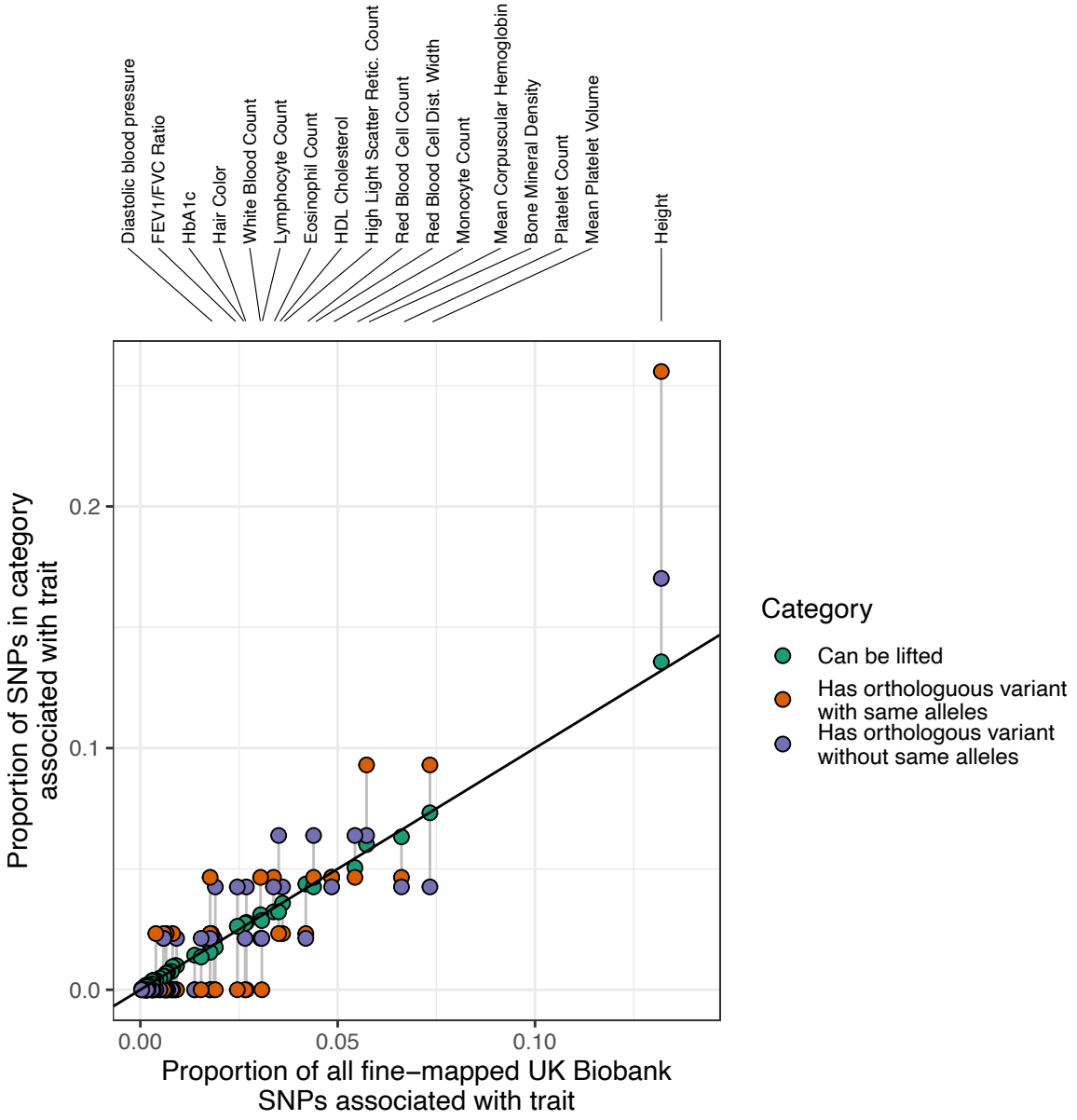


**Figure 5. Orthologues of fine-mapped variants linked to traits in the UK Biobank.** All 2240 fine-mapped UK biobank variants were lifted over to the cattle genome and the number that overlapped a variant in the cattle genome with and without the same alleles were determined. The Y-axis shows the proportion of all SNPs that could be lifted that are associated with each trait (green), that have an orthologue with the same alleles (orange) or that have an orthologue with different alleles (blue). These values are compared to the expected values (X-axis) as represented by the proportion of all fine-mapped variants linked to the given trait. Circles corresponding to the same trait are connected by vertical grey segments, with the trait indicated above for those traits with at least 55 fine-mapped variants. The proportion of all lifted variants that were associated with a given trait is strongly correlated with the original proportion of fine-mapped variants linked to that trait (green circles). However, variants with an orthologous cow variant are disproportionately associated with height, and in particular those with matching alleles in both species (orange circles). The data underlying this figure can be found in Supplementary Data 5.

### *Conservation of regulatory variation*

Most variants linked to important complex phenotypes are thought to be regulatory rather than coding[25]. To investigate whether regulatory variants are conserved across species we obtained the location of fine-mapped regulatory SNPs from the human GTEx[26] dataset. These human regulatory variants had been fine-mapped using three different approaches; CAVIAR[27], CaVEMaN[28] and DAP-G[29], and we took the superset of SNPs across all three. We then extracted the associations with orthologous genes of variants found at the orthologous location in cattle from the cattle GTEx project, who defined eQTLs across 23 different cattle tissues and cell types[30]. In total 221 of the human fine-mapped variants had a matching cattle variant in the cattle GTEx data that had been tested against the same orthologous gene in at least one tissue (Supplementary Data 6). Ignoring the allele change of the variants this number increases to 469. As shown in Figure 6a, these cattle variants at the orthologous position of the human fine-mapped variants are more likely to show an association (i.e. have a smaller p-value) than randomly sampled gene-variant pairs from the cattle GTEx cohort. This suggests these variants are often regulatory across both species. Notably this was largely observed whether restricting the cattle variants to those showing the same allele change as the human variant or not, with only a slight enrichment of smaller p-values among the former group in some tissues (Figure 6a). This suggests simply disrupting the same regulatory site may often be sufficient to affect the gene's expression across species in many cases.

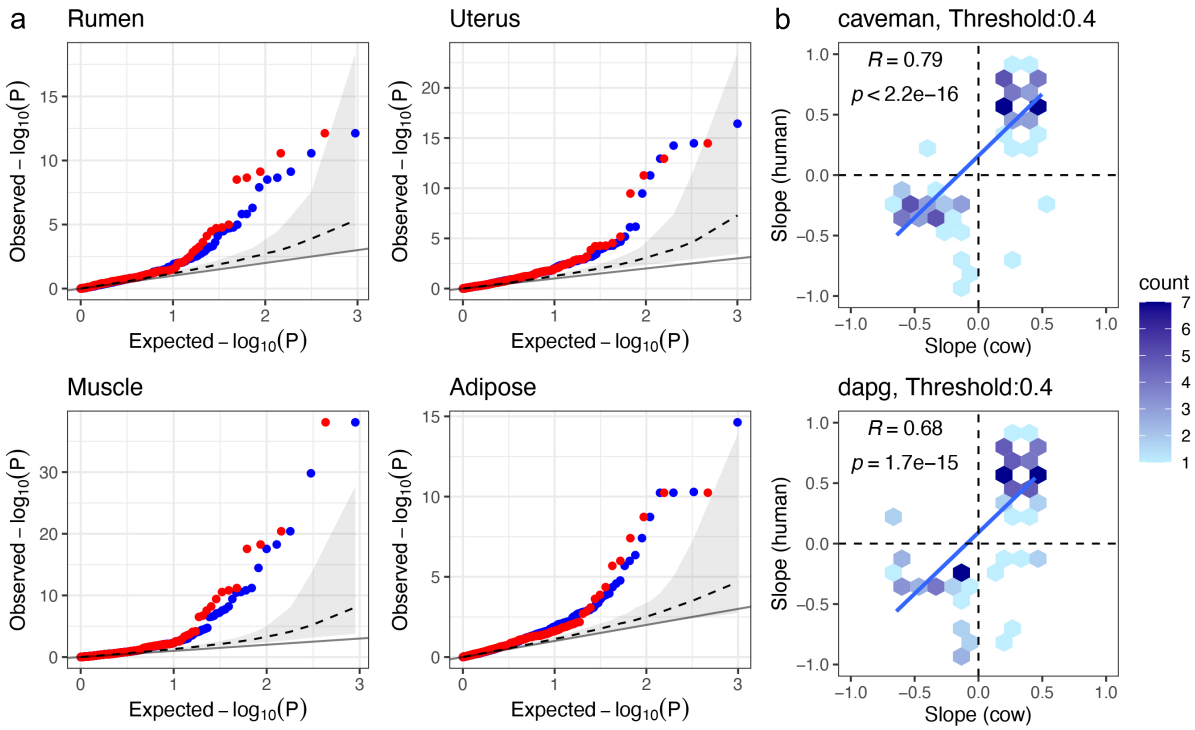



**Figure 6. Conserved effects of regulatory variants across humans and livestock.** (a) Quantile-quantile (Q-Q) plots of observed and expected cis-eQTL p-values of cattle variants that are direct orthologues of human fine-mapped regulatory variants. The blue points represent the observed and expected p-values of the cattle variant's association with the expression level of the cattle orthologue of the corresponding human gene in four cattle tissues irrespective of its allele change. The red points are the same after restricting to these variants exhibiting the same allele change as observed at the human SNP. The grey dashed line and grey ribbon represents the median and 95% confidence interval obtained when randomly sampling the same number of variants as shown by the blue points from all cattle variants tested (irrespective if have a known human orthologous variant or not) 1000 times. The line of parity (solid black) is also shown. This illustrates cattle orthologues of human fine-mapped regulatory variants are more likely to show evidence of also being linked to the orthologous gene's expression across different tissues. (b) Comparison of the slopes (direction of effects) of eVariants across species. The slopes of human fine-mapped regulatory variants (using caveman or dap-g approach) were compared to the slopes observed for their orthologues if they also had a significant cattle GTEx association with the expression level of the orthologous gene in cattle. The slope represents the impact on the gene's expression of increasing the dosage of the same allele in both species. Note the same eVariant can be found in multiple tissues and can therefore be represented multiple times in this plot. In total there are 83 and 106 human-cattle association pairs in the caveman and dapg plots, involving 43 and 57 distinct human eVariant-gene-tissue associations. The significant positive correlation remains if only one entry for each human eVariant-gene-tissue association is retained. This agreement in direction is seen despite not restricting to comparing the effects to the same tissues across species, i.e. the direction of effect is generally conserved across tissues as well as species. The data underlying this figure can be found in Supplementary Data 6.

The direction of effect of conserved regulatory variants were more likely to be conserved across the two species (Figure 6b), i.e. the same alleles are associated with increased or

decreased expression across species. This confirms that the effect of predicted functional variants appear often conserved. This is despite the different linkage disequilibrium patterns between the species, that may be expected to disrupt any conservation of direction of effects if these variants were not functional.

Figure 7 shows examples of the colocalization of eQTLs across humans and cattle. rs115287948 is a missense variant in the *SIRPA* gene that was fine-mapped in the GTEx cohort as a causative regulatory variant (probability > 0.5) linked to the expression of *SIRPB1* across a range of tissues including cultured fibroblasts and muscle (Figures 7a,b,c). A direct orthologue of this variant is also found at a co-localised eQTL in cattle muscle (Figure 7d) displaying an association with the same gene with the same direction of effect.

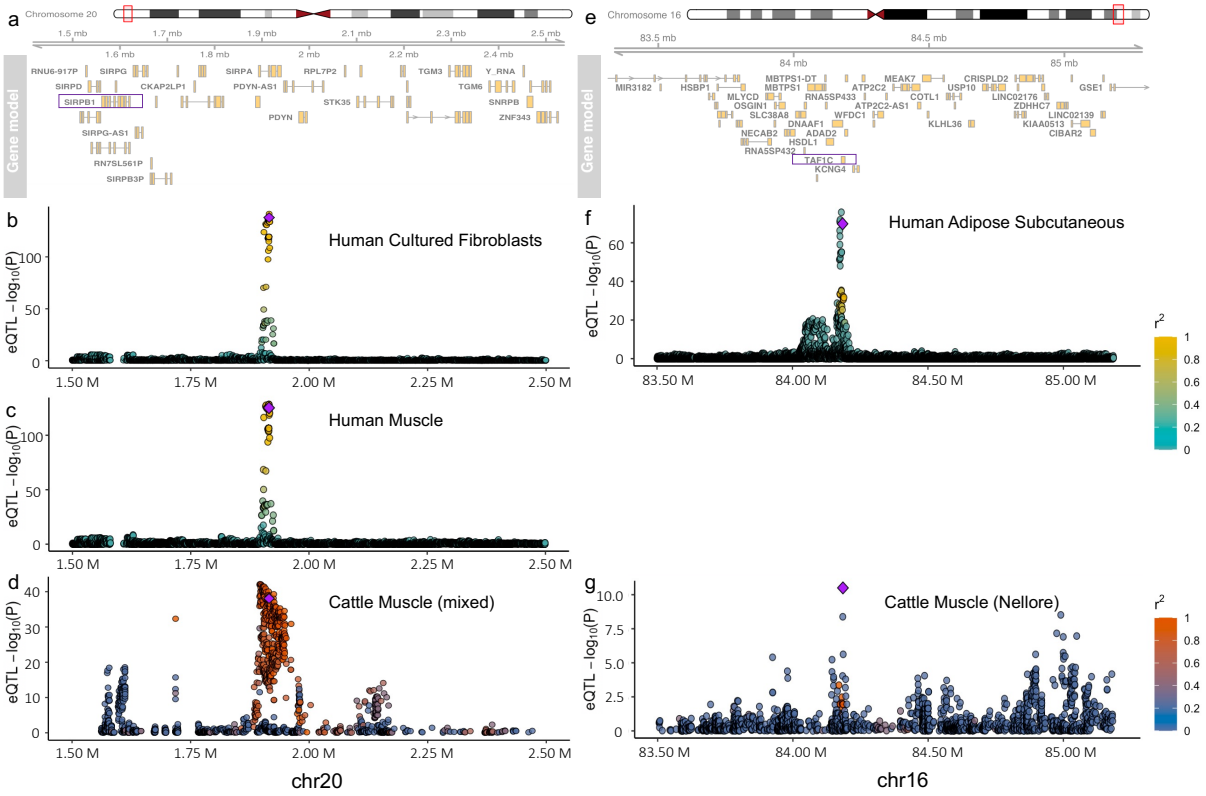



**Figure 7. Colocalization of eQTLs across humans and cattle.** (a) The human gene neighbourhood (hg38) of a shared eQTL, rs115287948, found across both humans and cattle. The gene regulated by the eQTL is indicated by a purple rectangle. (b) Strength of association of human variants with *SIRPB1* expression levels in cultured fibroblasts tissue. The fine-mapped regulatory variant, rs115287948, with a cattle orthologue is represented by the purple diamond.

Other variants are coloured according to their linkage disequilibrium ($r^2$) with this variant. (c) Strength of association of the same human variants but in muscle tissue. (d) Strength of association of variants with *SIRPB1* in cattle muscle tissue (mixed breeds). Each variant is plotted according to their orthologous position in the human genome and the variant with a fine-mapped orthologue in the human GTEx data is represented by the purple diamond. (e), (f), (g), same as (a), (b), (d) but for variant rs2230126 linked to the expression of *TAF1C* in different tissues. The data underlying this figure can be found in Supplementary Data 7.

rs2230126 is a variant falling within an alternative promoter of *TAF1C* with which it is a fine-mapped human regulatory variant (probability > 0.95) in a range of tissues including subcutaneous adipose (Figures 7e,f). An orthologue of this variant is also found in cattle and is the lead eVariant for the same gene in Nellore muscle tissue (Figure 7g).

We explored why some variants conserved in both species may not show evidence of impacting gene expression in cattle. Figure 8 illustrates predictions from the Enformer human deep learning model[31], that predicts transcriptional potential and chromatin states from DNA sequence alone. As shown in Figures 8a,b, Enformer predicts that the alternate allele of the rs10849334 variant is associated with reduced expression, specifically of an alternate, shorter isoform of the *NINJ2* gene. The predicted transcriptional potential of the TSS of longer isoforms of the *NINJ2* gene are unaffected by this variant. When the orthologous cattle DNA segment is run through Enformer the predicted transcriptional potential of the TSS at these longer isoforms remains, but the CAGE peak at the promoter of the shorter isoform is completely abrogated. Consistent with this we could also not find any evidence of a cattle orthologue of this shorter isoform in the public databases. Consequently, the lack of evidence of the cattle orthologue of rs10849334 affecting the expression of *NINJ2* is potentially due to the human variant specifically regulating this shorter isoform, that is absent in cattle due to sequence divergence in this locus.

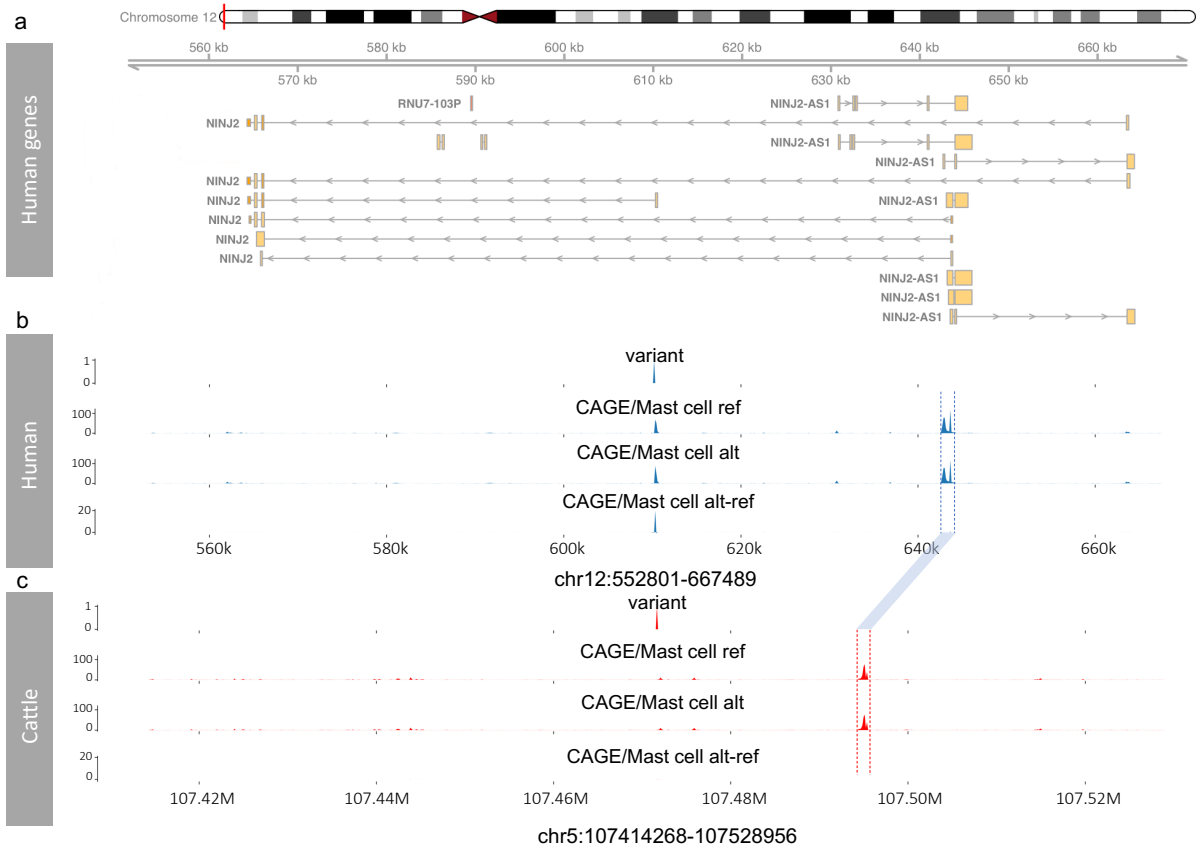



**Figure 8. Enformer predictions for human and cattle.** (a) The human gene neighbourhood of variant rs10849334. Different isoforms of the genes are included in the plot. (b) Enformer predicted Cap Analysis Gene Expression (CAGE) tracks of variant rs10849334 in mast cells (the track showing the largest human alt-ref difference). The first three tracks show i) the position of the target variant in the human genome, and the predicted CAGE levels for the ii) reference and iii) alternative alleles from Enformer, i.e., the predicted CAGE tracks obtained by taking the human DNA sequences containing reference and alternative alleles at the variant position as the inputs of the Enformer model. The final track shows the predicted difference between the CAGE levels from the reference and alternate sequences (alt-ref). As can be seen the only difference is observed specifically at the start position of the shorter *NINJ2* isoform in the centre of the plot. (c) Corresponding predicted CAGE tracks derived from the cattle sequences around the orthologous variant of rs10849334. The orthologous peaks at the TSSs of the longer *NINJ2* transcripts in cattle and human are indicated by the blue linking bar. However, no CAGE peak is predicted at the TSS of the shorter isoform of *NINJ2*, unlike with the human sequences. The data underlying this figure can be found in Supplementary Data 8.

In summary, although the effect of some variants do not lift over across species, potentially due to, for example, changes in the usage of isoforms between species, a range of human regulatory variants have orthologues in cattle that often have conserved effects and can consequently be used to provide insights into their mechanisms of gene regulation.

## Discussion

In this study we have demonstrated how millions of orthologues of human variants exist in domesticated species, including hundreds of orthologues of fine-mapped functional variants linked to diseases and phenotypes. These are consequently readily accessible large animal models of important human variants, that can potentially be studied at scale without the time and costs associated with transgenic approaches. Importantly we show that orthologous regulatory variants most often have conserved directions of effect across humans and cattle, suggesting their downstream effects can be effectively studied in these species.

Variants shared across humans and domesticated species are not restricted to one type of trait but are linked to a wide spectrum of phenotypes. From rare, monogenic disorders such as cystic fibrosis to highly polygenic traits such as height. However, variants associated with particular phenotypes are found across species more often than expected. For example, the 4 out of 23 variants associated with biotinidase deficiency that have a direct orthologue in cattle. It is in unlikely the co-occurrence of these variants is purely due to, for example, a higher mutation rate around the gene linked to this phenotype, as none of the other three species studied carry even one orthologue of these variants. This suggests there is a preferential overlap of variants linked to specific phenotypes, including more polygenic phenotypes such as height. Consequently, although the sharing of the overwhelming majority of variants across species is likely the result of neutral processes, the disproportionate sharing of variants linked to certain phenotypes potentially reflects selection to preferentially maintain such variants arising in each species. This not only provides insights into the evolution of these species, but also potential candidate variants for livestock breeding programs. The increased number of human height associated variants with an orthologue in cattle likely reflects the selection for body size in domesticated animals. However, as these variants remain polymorphic, the selective sweep is incomplete, and they remain suitable targets for breeding programs. Although there would be potential ethical concerns of introducing human variants into livestock species to improve their production, the same is not true if the variant already exists in

the species, as even editing the variant into another breed, would no longer come with the restrictions imposed on transgenic projects. Consequently, exploiting the large amount of data and studies on functional variants in humans, could potentially be leveraged to prioritise variants for testing their effects in livestock.

Despite the hundreds of orthologues of functional variants identified, this number is likely a substantial under-estimate of the true number of shared functional variants. This is because for a variant to be tied to a phenotype it needs to not only be at a sufficiently high frequency in the population to be discovered, but also with suitable data and patterns of linkage disequilibrium to be fine-mapped. However, most functional variants are, by their nature one or more of: rare, non-coding or in regions of elevated LD, and are therefore difficult to tie to a trait. Looking at the effects of rare variants across species may though help increase the pool of individuals in which to study their potential role. Likewise looking across species has the advantage that allele frequencies and linkage disequilibrium patterns can differ substantially and may, therefore, help in fine-mapping approaches. Extending this further, such approaches may help validate fine-mapping methods, for example by characterising which fine-mapping approach better identifies variants whose impacts are subsequently shown to be conserved across species. This is illustrated by the comparison of fine-mapped regulatory variants, and their conserved direction of effect across species. Of the three fine-mapping approaches studied, CAVIAR fine-mapped variants showed the lowest conservation of effect direction. This may reflect where variants are not truly functional, with the different patterns of LD in the different species with the actual causative variant meaning their eQTL coefficients are less conserved.

A caveat to such cross-species comparisons of regulatory variants is that not only can it be difficult to directly match tissues between species, but that power also generally differs due to differences in sample sizes. Consequently, the fact that a variant doesn’t also show evidence of being linked to a gene’s expression in another species doesn’t mean it is not a functional variant in both. Those variants we detect as being linked to gene regulation across species are likely those that are regulatory in multiple tissues, increasing the probability of us detecting it’s association in at least one.

It is likely that few if any human polymorphisms with orthologues in these mammalian species arose prior to the divergence of the respective species, and to still be polymorphic down the independent lineages. Rather they will have largely arisen independently in each. This is supported by the fact that despite there being millions of orthologues of human variants segregating in other mammals, few are found in more than one other species. Only twelve sites were polymorphic across the five studied mammals. Shared variants are simply most often found at sites with the highest mutation rates and lowest levels of purifying selection, and therefore reflect the increased chance of these sites mutating and not being purged from the population in both species. This indicates that the normalised presence or absence of orthologous variants may provide an alternate metric of the selective pressure on genomic regions, as illustrated by the depletion of orthologues of Clinvar variants across species.

Consequently the study of orthologues of human functional variants can be used across a range of studies. From understanding the biological mechanisms linking variants to important downstream phenotypes, to providing potential targets for livestock breeding and genome editing programs as well as understanding the selection pressures on our species.

## Methods

### *Genotype Datasets*

Previously published and filtered genotype data for five different species was used in this study. The genome-wide set of 78 million human SNPs from 2,504 individuals was obtained from the 1000 genomes consortium[15]. The dog genotypes from 722 individuals were obtained from Plassais et al[17]. The cattle and water buffalo genotypes of 477 and 79 individuals, respectively, were obtained from Dutta et al[9.], and the pig genotypes from across 409 individuals from the Genome Variation Map website[16]. All cohorts were subsequently restricted to biallelic SNPs only (cattle: 87,964,998; pig: 90,901,469; water

buffalo: 37,682,631; dog: 73,906,017). For all sets of human variants, their positions were lifted to their orthologous positions in the pig (SusScr3), cattle (BosTau9) and dog (CanFam3) genomes using the UCSC liftover utility[32] with chain files available from the UCSC website. For the water buffalo[33], where no public chain file exists, we used the nf-LO pipeline[34] to perform the liftover. Sites that were lifted to more than one location were excluded. SNPs from other species were said to have the same allele change as human SNPs if found at the orthologous position with alleles that directly matched or that matched their complement. This therefore assumes the ancestral base in these conserved regions is the same across mammals.

To test the impact of relatedness on the number of orthologous variants found in cattle, we used the relatedness2[35] parameter in vcftools[36] to identify pairs of animals with a kinship coefficient greater than 0. Individuals in each pair were then iteratively removed till the kinship coefficient between all pairs of remaining animals was 0 or less.

***Clinvar and UK biobank analyses***

The location of variants potentially linked to human health were downloaded from Clinvar[37], which contains SNPs linked to different human clinical phenotypes. Restricting this set to those labelled as “pathogenic” or “likely pathogenic” left 89,158 SNPs with likely functional consequences. Potentially functional variants linked to polygenic traits were obtained from Weissbrod et al.[22]. This study produced a list of 3281 fine-mapped, potentially functional variants associated with 47 complex traits of which 2240 were SNPs. The locations of these sets of SNPs were intersected with those from other species to identify those segregating in other mammals as described above. To test whether Clinvar variants linked to particular phenotypes are more likely to segregate in another species than expected, we used a Chi-squared test to examine whether the proportion of successfully lifted Clinvar variants linked to a particular phenotype that overlapped a variant with matching alleles was significantly higher than that observed across all other phenotypes. To test whether UK biobank variants lifted to the cattle genome were disproportionately associated with height a Fisher’s exact test was used, comparing the

proportion of variants with an orthologous variant with the same alleles that were linked to human height versus the proportion of successfully lifted variants linked to the same trait. To examine the impact of these orthologous variants on genes in humans, pigs and cattle, the variants were annotated using the Ensembl REST API in R, recording just the most severe reported consequence in each case.

***Regulatory variant analyses***

The GTEx v8 fine-mapped results for CaVEMaN, DAP-G and CAVIAR were downloaded from the GTEx portal. Together these reported 5,341,519 distinct tissue-gene-variant associations of which 2,145,167 could be lifted to an orthologous position in the cow genome. Upon filtering out variants that did not have a minimum probability > 0.2 in at least one of the datasets, this number reduced to 230,991 associations. These associations were then intersected with the cattle GTEx data to identify where an orthologous variant was significantly associated with the corresponding orthologous gene. Nominal p-values were obtained as described in the original cattle GTEx paper[30] and the human-cow gene orthologues were obtained from Ensembl version 103[38].

To examine whether cattle orthologues of human fine-mapped eVariants were more likely to show evidence of also being significantly associated with the expression level of the same gene, we extracted their corresponding p-values from the cattle GTEx data by tissue. The distribution of these p-values were then compared to the distributions of p-values for the same tissue of the same number of variants sampled from the total cattle GTEx data 1000 times to produce the shaded confidence intervals in the Q-Q plots.

To conservatively estimate false discovery rates for the cattle eQTLs we used the same random samples. For each real eQTL p-value we divided the average number of tissue-specific p-values across the 1000 samples that had a p-value as small or smaller by the corresponding number within the variants that were orthologues of human fine-mapped regulatory variants. This therefore corresponds to the approximate probability of having sampled a p-value as small or smaller from the background list of all variants tested in

the cattle GTEx project. This is conservative as a large number of the variants in this background list are eVariants. Therefore, this FDR corresponds to the false discovery rate above and beyond that expected given the number of regulatory variants in the background, and variants with a large FDR may still be regulatory variants.

To investigate why some regulatory variants shared across human and cattle may not have conserved impact on gene expression in cattle, we used Enformer, a deep learning architecture designed for predicting how DNA sequence influences gene expression[31]. We loaded the trained Enformer model, made predictions for reference and alternative alleles of each shared regulatory variant in both human and cattle, and obtained 5,313 predicted genomic tracks for each variant. The effect of each variant was evaluated by the difference between the reference and alternative predictions.

***Variant annotation and modelling***

The genome-wide set of 78 million human SNPs were annotated with 1589 features across four categories (Table 1), including sequence conservation, variants position properties, VEP[39] annotations and sequence context. For sequence conservation, we included 4 different conservation scores: phastCons100way, phastCons30way[40], phyloP100way and phyloP30way[41]. We downloaded bigWig files of these conservation scores from the UCSC genome annotation database[42] (hg38) and extracted the values at given positions using the pyBigWig python package[43]. To fully capture the position characteristics of the variants, we calculated the distance between the variants and different genome elements. We obtained the location of CpG islands from the UCSC genome annotation database[42], chromatin data (such as histone marks), TSS and regulatory features (enhancer, promoter, CTCF binding site and TF binding site) from Ensembl[38] (version 103). We used bedtools[44] closest command to calculate distances to CpG islands and chromatin data. The ChIPpeakAnno[45] R package was used for getting distances to the nearest TSS by biotype (only common biotypes were included, count >= 1000) and distances to various regulatory features. Then we used the VEP[39] command line tool to annotate the variants and get the allele frequencies and consequences of the

variants. Instead of using Reference/Alternative alleles, we used Ancestral/Derived alleles for the allele change. The human ancestral genome (hg38) was downloaded from Ensembl[38] (version 103) and the bedtools[44] getfasta command was used to extract the ancestral base. To get the 5-mer flanking sequences centered on the target variants, we used the samtools[46] faidx command. Then we calculated gene density (per megabase) of each variant using the findOverlaps and queryHits functions in the GenomicRanges package[47].

Using these genomic annotations as classification features, we trained machine learning models to predict whether a human variant has an orthologue in other livestock cohorts. Variants that have cattle orthologues (with matching alleles) were used as the foreground data in the models while variants without orthologues, i.e., variants that can be lifted to the cow genome but no cattle polymorphism was found, were used as background data. Similarly, we got foreground and background datasets for pig, water buffalo, the intersection of variants found across the cattle and pig cohorts, and the cross-species cohort (cattle, pig, dog and water buffalo). For the cross-species cohorts, variants with orthologues in any of the tested livestock species were used as foreground data and variants without orthologues in all tested species were used as background data. To avoid class imbalance problems, we down-sampled all background datasets to the same sizes as the foreground datasets for all models.

The feature tables were pre-processed before being used for model training. Data with missing values (found for sequence conservation scores) were discarded as they only accounted for a small proportion (1.4%) of the whole dataset. Categorical features, i.e., chromosome, consequence, allele change and 5-mer flanking sequences were encoded using different encoding methods (see Table 1). To minimize the introduction of new feature columns into the feature table and make the encoding more meaningful, a self-defined binary encoding method was used for sequence context features. We defined a dictionary for 4 bases (A: 1000, C: 0100, G: 0010, T: 0001) and mapped each base in the sequences to the corresponding binary string. The final strings for the sequences were split into binary columns and replaced the original categorical features in the feature table.

We constructed three tree-based machine learning models, Random Forest[48], XGBoost[49] and CatBoost[50] using the Scikit-learn Python package[48]. Models were trained on Eddie[51], a compute cluster of the University of Edinburgh, and 2 64GB GPUs on Eddie were used to train the CatBoost models. To enable balanced comparisons, subsets (200,000 data in total, 100,000 of which was foreground data and 100,000 background data) of the datasets for different species were used. Each subset was divided into a training set and test set at the ratio of 70% and 30%. We used 5-fold cross-validation to evaluate our models on the training sets. To improve the performance of the models, we used random search[52] and manual tuning methods for hyper-parameter tuning.

***Statistics and reproducibility***

Comparisons between two groups were conducted via Chi-Squared test, Two-sample Kolmogorov-Smirnov test and Fisher's exact test as indicated in the paper. All statistical analyses were performed using R.

**Data availability**

The human 1000 genomes cohort genetic variants were obtained from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/, the pig genotypes from ftp://download.big.ac.cn/GVM/Sus_scrofa/SNP/detailed_vcf/all_SNP.vcf.gz, the dog genotypes from https://sra-pub-src-1.s3.amazonaws.com/SRZ189891/722g.990.SNP.INDEL.chrAll.vcf.1 and the cattle and water buffalo genotypes were those published in [9]. The set of annotated variants used in the machine learning analyses can be found at https://doi.org/10.6084/m9.figshare.20401851.v1. Source data underlying Figure 1 is provided in Supplementary Data 1. All other data is available upon reasonable request.

**Competing interests**

The authors declare that there are no competing interests.

**Author contributions**

J.P. conceived the initial project idea, further developing it with R.Z., A.Ta. and M.H.. R.Z. and J.P. performed the majority of analyses with contributions from A.Ta., L.F., S.L., G.L., N.C.H. and A.Te.. L.F., S.L., G.L. and A.Te generated the cattle GTEx data used in this study. J.P. and R.Z. wrote the initial manuscript draft with all authors contributing to the final version.

**Acknowledgements**

This work was supported by grant BB/W000288/1 and Institute Strategic Programme Grant BBS/E/D/10002070 from the Biotechnology and Biological Sciences Research Council (BBSRC).

## References

1. Käser, T. Swine as biomedical animal model for T-cell research—Success and potential for transmittable and non-transmittable human diseases. *Mol. Immunol.* **135**, 95–115 (2021).
2. Meurens, F., Summerfield, A., Nauwynck, H., Saif, L. & Gerdts, V. The pig: a model for human infectious diseases. *Trends Microbiol.* **20**, 50–57 (2012).
3. Ziegler, A., Gonzalez, L. & Blikslager, A. Large Animal Models: The Key to Translational Discovery in Digestive Disease Research. *Cell. Mol. Gastroenterol. Hepatol.* **2**, 716–724 (2016).
4. Walters, E. M. & Prather, R. S. Advancing Swine Models for Human Health and Diseases. *Mo. Med.* **110**, 212–215 (2013).
5. Sjöstedt, E. *et al.* An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* **367**, eaay5947 (2020).
6. Lunney, J. K. *et al.* Importance of the pig as a human biomedical model. *Sci. Transl. Med.* **13**, eabd5758 (2021).
7. Zhu, F., Nair, R. R., Fisher, E. M. C. & Cunningham, T. J. Humanising the mouse genome piece by piece. *Nat. Commun.* **10**, 1845 (2019).

8. Hayes, B. J. & Daetwyler, H. D. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annu. Rev. Anim. Biosci.* **7**, 89–102 (2019).
9. Dutta, P. *et al.* Whole genome analysis of water buffalo and global cattle breeds highlights convergent signatures of domestication. *Nat. Commun.* **11**, 4739 (2020).
10. Pir, M. S. *et al.* ConVarT: a search engine for matching human genetic variants with variants from non-human species. *Nucleic Acids Res.* **50**, D1172–D1178 (2022).
11. Broekema, R. V., Bakker, O. B. & Jonkers, I. H. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* **10**, 190221 (2020).
12. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
13. Bouwman, A. C. *et al.* Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat. Genet.* **50**, 362–367 (2018).
14. Raymond, B. *et al.* Using prior information from humans to prioritize genes and gene-associated variants for complex traits in livestock. *PLOS Genet.* **16**, e1008780 (2020).
15. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
16. Li, C. *et al.* Genome Variation Map: a worldwide collection of genome variations across multiple species. *Nucleic Acids Res.* **49**, D1186–D1191 (2021).
17. Plassais, J. *et al.* Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat. Commun.* **10**, 1489 (2019).
18. Fryxell, K. J. & Moon, W.-J. CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content. *Mol. Biol. Evol.* **22**, 650–658 (2005).

19. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
20. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
21. Lean, I. J. & Rabiee, A. R. Effect of feeding biotin on milk production and hoof health in lactating dairy cows: A quantitative assessment. *J. Dairy Sci.* **94**, 1465–1476 (2011).
22. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
23. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
24. Wilkin, D. J. *et al.* Mutations in fibroblast growth-factor receptor 3 in sporadic cases of achondroplasia occur exclusively on the paternally derived chromosome. *Am. J. Hum. Genet.* **63**, 711–716 (1998).
25. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020).
26. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
27. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics* **198**, 497–508 (2014).
28. Brown, A. A. *et al.* Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat. Genet.* **49**, 1747–1751 (2017).
29. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, e1006646 (2017).
30. Liu, S. *et al.* A comprehensive catalogue of regulatory variants in the cattle transcriptome. 2020.12.01.406280 (2021) doi:10.1101/2020.12.01.406280.

31. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
32. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590-598 (2006).
33. Low, W. Y. *et al.* Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat. Commun.* **10**, 260 (2019).
34. Talenti, A. & Prendergast, J. nf-LO: A Scalable, Containerized Workflow for Genome-to-Genome Lift Over. *Genome Biol. Evol.* **13**, evab183 (2021).
35. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
36. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
37. Landrum, M. J. & Kattman, B. L. ClinVar at five years: Delivering on the promise. *Hum. Mutat.* **39**, 1623–1630 (2018).
38. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).
39. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
40. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
41. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
42. Navarro Gonzalez, J. *et al.* The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* **49**, D1046–D1057 (2021).
43. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
44. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
45. Zhu, L. J. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, 237 (2010).

46. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
47. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
48. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
49. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016). doi:10.1145/2939672.2939785.
50. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. (2017).
51. *Edinburgh Compute and Data Facility web site*. (U of Edinburgh, 2021).
52. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).

## 3.3 Discussion

In this study, we conducted an extensive exploration into the occurrence of orthologues of human variants in domesticated species, including cattle, water buffalo, dog, and pig. Our results revealed that over 1.6 million human variants were observed in at least one of these four mammalian species. This number is expected to increase with larger animal cohorts. Among these, we identified hundreds of orthologues associated with fine-mapped functional variants linked to various diseases and traits, such as height. Furthermore, by investigating regulatory variants we discovered that both the regulatory effects and the direction of effects exhibit conservation across humans and cattle, suggesting that the exploration of their downstream consequences can be effectively conducted in these respective species. These readily accessible large animal models, possessing naturally occurring orthologues of variants, provide valuable opportunities for studying the biological impact of human variants at scale.

Using machine learning we explored the features associated with the probability of a human variant having a livestock orthologue. Among various machine learning models, conservation score, allele change and 5-mer flanking sequence consistently emerged as the top features. By further investigating the relationship between the conservation scores and the orthologous variants, we found that human variants located in less-conserved regions are more likely to possess a livestock orthologue. This is consistent with variants occurring in less-conserved regions not being under directional selection, where they would otherwise ultimately become fixed in one or the other species. Without strong selective pressure, these neutral variants are more likely to persist across different species. This observation is further supported by the finding that functional Clinvar variants, primarily located in conserved coding regions, are less likely to have a livestock orthologue.

Despite the presence of millions of human variants having an orthologue in at least one of the investigated mammalian species, the number of variants found across multiple species is notably smaller, and this count decreases to only twelve when considering conservation across all four species. It is probable that few if any human variants with

livestock orthologues emerged before the divergence of the respective species and remained polymorphic throughout their independent evolutionary lineages. Instead, these variants predominantly arose independently in each species, leading to the small number of variants found across multiple species. Furthermore, the differences in the number of shared variants among different pairs of animals can be attributed to multiple reasons. In addition to the differences in the numbers of variants within each species, this can also potentially be due to the orthologous regions between closer species being mapped better than those between distant species.

While many variants are shared across species because of neutral evolutionary processes, there is a notable enrichment of variants linked to specific phenotypes. For instance, human variants associated with biotinidase deficiency are disproportionately present in cattle (4 out of 23 variants), while none of them are conserved in the other three studied species. Several factors could explain this observation, including dietary differences and physiological differences. The typical diet of cattle may be richer in biotin compared to the human diet, which could compensate for reduced enzyme activity caused by the variants. Additionally, as ruminants, cattle possess a complex digestive system with a rumen microbiome that plays a significant role in nutrient breakdown and absorption (K. Liu et al., 2021). This might offer alternative pathways for biotin acquisition. Furthermore, this observation suggests that any selective pressure on these variants is potentially species-specific. The distinct genetic and evolutionary histories of different species result in varying selective pressures acting on specific genetic variants. Therefore, the conservation of these variants can differ among different species. In addition to offering valuable insights into the evolutionary paths, the enrichment of conservation of specific variants associated with particular traits also provides valuable insights into breeding values. For example, the observed increase in the number of human height-associated variants with cattle orthologues likely indicates the selection for body size traits in livestock species. The presence of these polymorphic variants in the population suggests that the selective sweep is incomplete, making them appropriate candidates for inclusion in breeding programs. The substantial studies on these trait-associated variants

in humans can potentially be leveraged to aid in prioritising orthologous variants with breeding values in livestock species if they're not deleterious.

Most functional variants possess characteristics such as association with regions of elevated linkage disequilibrium (LD), which makes it difficult to link specific variants in the region to the traits. Examining the effects of these functional variants across species can be advantageous in enlarging the cohort of individuals available for investigating potential functional variants. In the same way that genes linked to traits such as high-altitude adaptation have been better fine-mapped by looking across species (Witt & Huerta-Sánchez, 2019).

In the study, the conservation of the effects and the direction of effects of three sets of regulatory variants based on different fine-mapping approaches between human and cattle was investigated. Among these three approaches, regulatory variants based on the CAVIAR fine-mapping method exhibited the lowest conservation of effect direction. This may potentially be attributed to an enrichment of non-functional variants being identified with this approach.

A limitation of the study is that, as shown in the paper, the machine learning models did not achieve good performance on predicting human variants with or without livestock orthologues, with a highest AUC score being 0.73. This is primarily due to the inherent complexity of the task, which involves modelling a process largely driven by randomness, i.e., the co-occurrence of the sites of random mutations in two different species. While selection and factors such as CpG deamination make this process less random, they do not remove the underlying randomness of the fact that any base can mutate in a given generation. Therefore, it can be challenging to identify orthologous variants between these species using machine learning models. There are several potential approaches worth considering for further exploration. First, other machine learning algorithms such as LightGBM can be explored as alternatives to improve the prediction accuracy. Additionally, while the models employed in the study are already ensemble models, further combination of these models may potentially enhance their overall performance. Another approach is to optimise the probability threshold for determining the predictions, which is

set to 0.5 by default. Since our focus is on correctly identifying variants with livestock orthologues, increasing the threshold can help minimize false positives, i.e., reduce the instances of classifying human variant without livestock orthologue as having orthologue. However, this may also result in more missed true positives, i.e., human variants with livestock orthologues. Therefore, it is necessary to find a threshold that balances true positives and false positives.

There are some other potential improvements of the study. First, it may be meaningful to explore variants conserved across three species and compare their conservation rate with those of sharing across two species. Besides, an investigation of the functional consequences of these shared variants may provide insights into key biological pathways and mechanisms that have been preserved across multiple species throughout evolutionary time. Second, Halldorsson et al. recently introduced a novel cohort of human variants, including over half a billion SNPs (Halldorsson et al., 2022), which is much larger than the 1000 genomes cohort employed in our present study. With a larger cohort, the sample size is expanded, leading to an increased ability to find shared variants and facilitating more robust and comprehensive analyses of variant conservation across species. Furthermore, the inclusion of variants from a broader range of human populations within the larger cohort enhances the generalization of the findings.

# Chapter 4 Using machine learning to predict regulatory variants in human and cattle

## 4.1 Introduction

Regulatory variants play an important role in shaping the complex mechanisms that control gene expression, accounting for up to 60% of the genetic differences relevant to disease (Nica & Dermitzakis, 2013). Therefore, the identification of regulatory variants can help facilitate the interpretation of disease and trait causation in both human and livestock species, providing valuable guidance for downstream research, such as livestock genome editing or marker-assisted breeding. However, these regulatory variants are predominantly located in non-coding regions, which have remained relatively underexplored, consequently posing challenges for their accurate functional interpretation.

As introduced in Chapter 1, Genome-Wide Association Studies (GWAS) have identified genomic regions that are associated with diseases and other important traits. However, pinpointing the causal variants within these regions remains a challenge due to several factors, including their enrichment in non-coding regions and linkage disequilibrium (LD) among the variants. Expression Quantitative Trait Loci (eQTL) analysis can be conducted to identify variants that are linked to changes in gene expression levels (Nica & Dermitzakis, 2013). In eQTL analyses, statistical tests are employed to assess the association between variants and gene expression levels. However, similar to GWAS, eQTL analysis examines the correlation between variants and gene expression, which does not necessarily imply causality, for example due to the confounders of LD or uncontrolled covariates such as population stratification.

Statistical fine-mapping is an approach used to determine the most likely causal variants from neighbouring variants in LD in a locus from studies like GWAS or eQTL analysis (Maller et al., 2012). Various fine-mapping approaches have been developed, from the traditional heuristic methods (Schaid et al., 2018) to more accurate and widely adopted

Bayesian approaches (Hutchinson, Watson, et al., 2020). Furthermore, some research also integrates functional genomic annotations with the fine-mapping approach to improve the efficacy of identifying causal variants (Weissbrod et al., 2020). The principles underlying these fine-mapping approaches were discussed in detail in Chapter 1.1.3.4. Nevertheless, fine-mapping approaches have some limitations. The power of these approaches is influenced by sample size, with smaller sample sizes resulting in limited ability to differentiate closely linked variants. Additionally, variants of functional importance but with small effect sizes may be excluded from fine-mapping results.

Some state-of-the-art computational methods, such as machine learning, have been used to identify regulatory variants. Machine learning approaches are designed for finding patterns from large-scale data, making them particularly suitable for analyzing genomic data. These methods have found widespread utility in addressing genomic challenges, including predicting variant effects, gene expression, and regulation. The detailed application of machine learning approaches in predicting regulatory variants was discussed in Chapter 1.2.3.2. Machine learning algorithms can capture complex relationships between variant categories and their associated features, offering insights into the underlying characteristics linked to specific variants such as regulatory variants. Therefore, the prediction results obtained through machine learning can serve not only as a complement or support to traditional statistical fine-mapping strategies but also contribute to a deeper comprehension of the characteristics of causal regulatory variants.

Unlike the relatively abundant high-quality regulatory variant datasets available for human, livestock species face a pronounced scarcity of such data. Identifying regulatory variants that influence gene expression and subsequently impact downstream phenotypes is crucial for guiding research aimed at enhancing economically important or welfare-related traits in livestock species. Additionally, the current use of machine learning to predict variant effects has predominantly focused on humans. A primary limiting factor for the application of machine learning in livestock species is the lack of important variant annotations suitable for machine learning in these less well-annotated species, such as conservation scores. The variant annotation pipeline proposed in Chapter 2 has

expanded the annotations to livestock species, providing an opportunity to extend these machine learning-based approaches to a broader range of species, thereby aiding in the prioritization of regulatory variants in less extensively studied species.

### 4.1.1 Regulatory variant datasets in human and cattle

The pivotal aspect of applying machine learning approaches lies in the selection of data used for model training. The upper limit of the model's performance is typically constrained by the training data itself, while the choice of machine learning algorithms and subsequent model tuning will generally have a comparatively limited influence on model performance. The availability of novel regulatory datasets has opened the possibility of employing machine learning approaches in this domain. In this section, I will introduce the potentially accessible datasets that can be used for machine learning model training and testing in both human and cattle.

The Genotype-Tissue Expression project (GTEx) has focused on how genetic variation impacts gene expression and consequently contribute to phenotypic diversity in human (GTEx Consortium, 2020). In GTEx, 49 tissues/cell lines were included, each meeting the criterion of having at least 70 individuals with available RNA sequencing (RNA-seq) and genotype data derived from whole-genome sequencing (WGS). This aggregation yielded a comprehensive dataset comprising 15,201 samples contributed by 838 donors from diverse populations including European American, African American, Asian American, and Hispanic/Latino. Among the initial 43,066,422 variants, 4,278,636 variants were identified as significantly associated with gene expression in at least one tissue.

Subsequently, they employed three different fine-mapping approaches: CAVIAR, DAP-G and CaVEMaN, to deduce causal regulatory variants in each locus across different tissues (GTEx Consortium, 2020). Both the CAVIAR and DAP-G approaches rely on the Bayesian fine-mapping method, sharing the assumption that each risk locus may encompass multiple casual variants (Hutchinson, Asimit, et al., 2020). These two approaches utilize different search strategies in the region, with DAP-G being capable of incorporating functional annotations that are not available in CAVIAR. On the other hand,

the CaVEMaN approach employs a frequentist definition of causal probability instead of the Bayesian method (A. A. Brown et al., 2017). Another notable feature of CaVEMaN is that, instead of assuming one or more variants as potential causal variants in each locus, it utilizes a regression approach to estimate the number of causal eQTLs and then map each eQTL independently. In the CaVEMaN paper, a comparison was conducted among these three approaches, revealing a general agreement in terms of causal probability. Notably, both CAVIAR and DAP-G exhibited underestimation issues when compared to CaVEMaN.

Following the human GTEx project, the Farm Animal Genotype-Tissue Expression (FarmGTEx) project started the exploration of regulatory variants in farm animals, such as cattle and pigs (S. Liu et al., 2022; Teng et al., 2024). In the cattle GTEx project, an analysis was conducted on 7,180 publicly available RNA-Seq samples originating from 46 cattle breeds and breed combinations, encompassing 114 distinct tissues (S. Liu et al., 2022). From all these samples, a median of 21,623 Single Nucleotide Polymorphisms (SNPs) were identified. Through subsequent imputation, the SNP dataset for each sample was expanded to include up to 3,824,444 SNPs. Downstream eQTL study focused on 23 distinct tissues, each with more than 40 individuals. Subsequently, DAP-G was used to identify causal variants within each genomic locus. However, while the cattle GTEx project serves as one of the most extensive reference resources for the cattle transcriptome available to date, it does have some limitations. Notably, the inclusion of a limited number of individuals and breeds in the research raises concerns about the imputation accuracy for underrepresented breeds. Additionally, unlike human GTEx, the data used in cattle GTEx was collected from disparate studies. Furthermore, factors such as a minority of tested variants and large disparities in sample sizes across different tissues are likely to influence the final outcomes of the fine-mapping process.

Apart from the eQTL data from cattle GTEx, our group has obtained regulatory variants through Massively Parallel Reporter Assays (MPRAs) strategy in cattle aortic endothelial cells. MPRA is a high-throughput experimental technique for testing the regulatory potential of DNA sequences, i.e., identifying regulatory elements (Mulvey et al., 2021),

and can be further applied to discover regulatory variants. The MPRA approach couples genomic features, such as individual alleles of a genomic sequence, with a reporter gene containing distinct transcribed barcodes. This enables the simultaneous and multiplexed measurement of element activity at the RNA level (Mulvey et al., 2021). MPRA offer numerous advantages, including efficiency and the avoidance of LD issues. However, it is important to note that MPRAs are often conducted within isolated and controlled cellular environments, potentially limiting their ability to fully capture the function of elements in their native genomic context. The MPRA strategy we are using is the Survey of Regulatory Elements (SuRE) technology, which was first employed in humans to successfully identify over 30,000 regulatory SNPs, also known as reporter assay QTLs (raQTLs) (van Arensbergen et al., 2019).

These datasets serve as valuable sources for downstream machine learning applications in both humans and cattle, offering opportunities to delve deeper into the underlying characteristics of regulatory variants and to make predictions for new ones.

### 4.1.2 Regulatory variant location validation

As discussed earlier, regulatory variants primarily reside in non-coding regions, particularly within or near regulatory elements. To validate the location of the predicted regulatory variants, datasets containing regulatory elements obtained from several techniques can be utilized. Precision Run-On Capturing (PRO-cap) is a technique that enables the precise mapping of RNA polymerases actively engaged in transcription across the genome, thus facilitating the identification of transcription start sites (TSS) (Mahat et al., 2016). Moreover, PRO-cap possesses the ability to capture TSS during the early stages of RNA synthesis, distinguishing itself from alternative TSS analyses that rely on mature RNA, thus offers a distinct benefit in the identification of enhancers. An alternative technique for TSS identification is Cap Analysis of Gene Expression (CAGE). In this approach, a modified cap structure is added to the 5’end of nascent RNA molecules, serving as a marker for TSS identification (Kodzius et al., 2006). In a recent study, CAGE was employed to pinpoint TSS along with their co-expressed short-range enhancers (within 1kb) across 24 distinct tissues from three cattle populations: dairy, beef-dairy cross,

and Canadian Kinsella composite cattle (Salavati et al., 2023). These datasets provide opportunities to validate the locations of predicted regulatory variants in the genome.

### 4.1.3 Objectives

The objective of this chapter was to utilize the annotation workflow from Chapter 2 to train machine learning models to predict regulatory variants in both humans and cattle. The performance of various machine learning algorithms was compared, models trained using different datasets were evaluated, and the inherent characteristics associated with a variant being predicted as a regulatory variant was explored. Furthermore, different strategies were employed and compared in cattle regulatory variant prediction, encompassing methods exploiting human annotations as well as those based purely on cattle annotations.

## 4.2 Materials and methods

In this section, I start by introducing the data preparation for both human and cattle, and how these data were processed to ensure compatibility with downstream machine learning models. Following this, I provide a comprehensive description of the construction, improvement, and evaluation of the machine learning models.

### 4.2.1 Human data preparation

#### 4.2.1.1 Human regulatory variants and annotations

Human fine-mapped regulatory variants were obtained from the CaVEMaN dataset from GTEx (GTEx Consortium, 2020). The variants at cis-eQTLs included in the CaVEMaN dataset were defined as foreground variants, i.e., the target regulatory variants to be predicted. The remaining variants in the GTEx reference file, comprising all tested variants, were considered as background variants. Prior to any processing procedures, there were 1,314,543 foreground variants fine-mapped in at least one of the 49 different tissues and 45,255,160 background variants.

As the CaVEMaN foreground variants were restricted to +/- 1Mb around the TSS, the background variants were also restricted to the same range to maintain consistency. with the foreground data. Furthermore, sex chromosomes were removed from the dataset because they possess unique biological characteristics compared to autosomes, which could potentially introduce confusion in the downstream model. Subsequently, duplicate variants in the datasets were removed, and only SNPs were retained for analysis. The background dataset was randomly down-sampled to match the number of variants in the foreground data, ensuring class balance. As a result, there were 590,778 variants left in each class. The summary of different human variant sets used in the modelling work can be found in Figure S4.3 (A).

Subsequently, these 1,181,556 variants were annotated with 1,583 annotations using the variant annotation pipeline presented in Chapter 2. It should be noted that the allele frequency annotations were not used as features in the subsequent studies. This is because, the association between a regulatory variant and a gene cannot be detected unless the variant has a high enough allele frequency, leading to an artificial distribution difference of the allele frequency in foreground and background data. This discrepancy is more statistical than biological, and it could potentially mislead the model and result in an over-evaluation of the model's performance. The Enformer sub-workflow was employed to annotate only high-confidence regulatory variants, i.e., those possessing a CaVEMaN causal probability exceeding 0.5. The constraints posed by limited GPU resources on Eddie (University of Edinburgh high performance computing system) rendered it difficult to annotate the entire dataset with the high-dimensional Enformer features.

#### 4.2.1.2 Additional annotations based on EpiMap

In addition to the annotations in the annotation pipeline, I extracted a complementary set of features from EpiMap, a comprehensive compendium that incorporates over 17,000 epigenomic data in humans (Boix et al., 2021). EpiMap comprises datasets based on 18 different marks/assays, such as ATAC-seq, DNase-seq and H3K4me1, spanning multiple tissues. The BigWig files for 593 averaged tracks per group (tissue and mark) were downloaded from https://epigenome.wustl.edu/epimap/data/, and then the scores for

each variant was extracted from the track files using the pyBigWig (Ramírez et al., 2016a) package in Python.

### 4.2.2 Cattle data preparation

For cattle data preparation, two sets of regulatory variants were included, those at regulatory variants from the cattle GTEx dataset and the regulatory variants from the SuRE analysis.

#### 4.2.2.1 Cattle GTEx regulatory variants

In the cattle GTEx dataset, each variant had been assigned two p-values: one nominal p-value, which indicates the significance of the association between a variant and the expression levels of a target gene, and the other one is the permutation p-value, a corrected version of the nominal p-value that takes into account multiple testing. Initially, the variants with the smallest nominal p-value for each gene across all tissues were extracted (i.e. the lead variant). If there were more than one variant associated with each gene that possess the same smallest p-value, they were all included in the dataset. Then these variants were filtered based on their associated permutation p-values, with a maximum threshold of 0.05. Furthermore, only SNPs were retained in the dataset. After filtering, 79,215 variants were retained as foreground regulatory variants, and an equivalent number of background variants were randomly down-sampled from the variants with a nominal p-value greater than 0.05 in the remaining dataset, resulting in a total of 158,430 variants.

#### 4.2.2.2 MPRA regulatory variants

Another set of cattle regulatory variants was obtained using the SuRE approach. The SuRE analysis and downstream raQTL calling were conducted by Annogen, a company that specializes in MPRA. Aortic endothelial cells from two animals, *Bos indicus* and *Bos taurus*, were utilized in the assays. The SuRE signal for every allele at each SNP was computed by employing the normalized barcode expression value corresponding to each barcode within the SuRE libraries. Next, a two-sided Wilcoxon rank-sum test was executed for each SNP to generate the p-values and false discovery rate (FDR). The

initial variant set was filtered by genotype quality (GQ) using a threshold of 20 and removing variants with a sequencing depth (DP) of 4 standard deviations (SDs) above the mean. Subsequently, variants with uncalled alleles in matching 30x WGS data or those exhibiting homozygous genotypes were excluded from the dataset. The SuRE paper (van Arensbergen et al., 2019), defined regulatory variants as variants with a SuRE signal > 4 and a 5% false discovery rate (FDR). In my case, I utilized a more stringent SuRE signal threshold of 5 and a Wilcox p-value cut-off of 0.006685 on the remaining dataset after GQ and DP filtering. Finally, 18,782 regulatory variants were retained as foreground data, and the same number of background variants were randomly sampled from the remaining variants with p-value > 0.006685 and SuRE signal > 5. The summary of different cattle variant sets used in the modelling work can be found in Figure S4.3 (B).

#### 4.2.2.3 Cattle variant annotations

To compare two approaches for cattle regulatory variant prediction, i.e., one based on human annotations and the other one using cattle annotations, the positions of cattle variants were initially lifted over to the human genome (hg38). The UCSC liftOver command line tool (Kuhn et al., 2013) and the chain file bosTau9ToHg38.over.chain.gz from https://hgdownload.soe.ucsc.edu/goldenPath/bosTau9/liftOver/ were employed for lift-over. All 158,430 cattle GTEx variants were successfully lifted over from cattle to human. Subsequently, the lifted variants were annotated using human annotations, while the original cattle variants were annotated using cattle annotations. For human annotations, only the following were retained: allele change, distance to TSS, distance to CpG island, conservation scores, gene density, flanking sequence, and variant consequence. These annotations were selected because they are also available for cattle, enabling direct comparison of model performance between the two approaches.

### 4.2.3 Feature engineering

Feature engineering work is a crucial part in machine learning that involves the transformation of initial annotations into final features in a suitable format for downstream machine learning models. This section provided a detailed exposition on the feature engineering approaches utilized in the project. The primary package used in this section

and the subsequent machine learning section was Scikit-learn v0.24.2 (sklearn), a comprehensive machine learning package in Python (Pedregosa et al., 2011). Figure 4.1 depicts the workflow for the feature engineering process and the downstream machine learning steps.

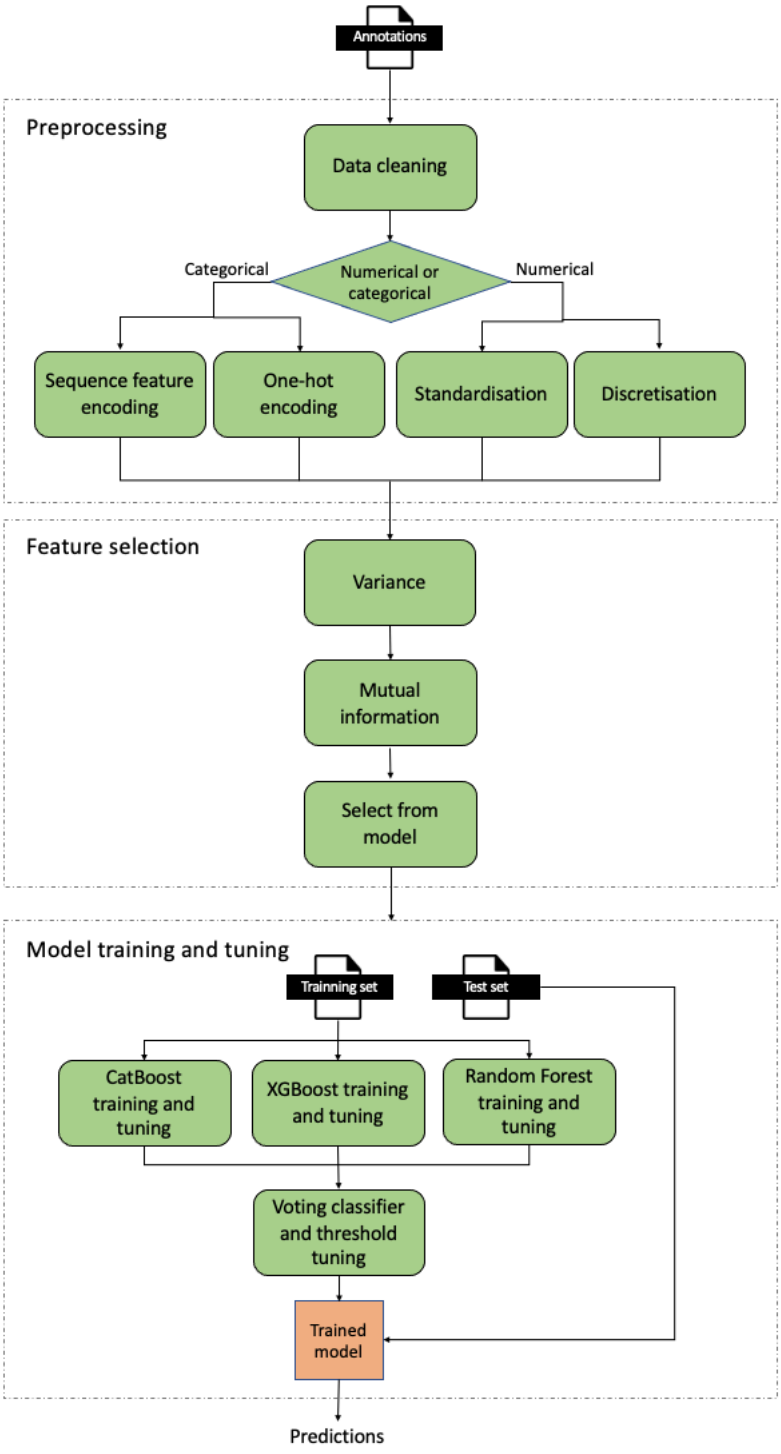


**Figure 4.1 The workflow for both the feature engineering process and the subsequent machine learning process.**

#### 4.2.3.1 Feature pre-processing

After obtaining the variant annotations for human and cattle, these initial annotations underwent pre-processing to remove any abnormal data, such as missing values, from the dataset. The Python package Pandas Profiling v2.11.0 (https://github.com/pandas-profiling/pandas-profiling) was employed for an initial check of the raw annotations. Missing values were identified in the conservation score annotations, accounting for approximately 0.9% of human variants and 0.8% of cattle variants. Given their small proportion within the dataset, instead of employing imputation strategies for the missing values, variants with missing conservation scores were simply excluded from the dataset. Subsequently, the numerical features within the annotations were standardised using the StandardScaler library in sklearn (Pedregosa et al., 2011), which involved subtracting the mean and scaling each feature to have unit variance.

#### 4.2.3.2 Feature encoding

In addition to the numerical features in the initial annotations, there were several categorical features that required encoding before being used as input for subsequent machine learning models. These categorical features encompassed allele change, chromosome name, flanking sequence, and variant consequence. They can be divided into two groups: the first group comprises features that exclusively represent discrete categories, including chromosome name and variant consequence; the second group comprises features that possess potential informative content, including allele change and flanking sequence. The first group of categorical features was subjected to the conventional one-hot encoding technique (Pedregosa et al., 2011). Meanwhile, the encoding of allele change and flanking sequence was accomplished through a custom-defined mapping strategy. This approach defined a dictionary for four different bases (A: 1000, C: 0100, G: 0010, T: 0001), and then each base within the feature was mapped to the corresponding string representation. The resultant string, including binary values of 0

and 1, was further partitioned into separate integer columns, each indicating the presence or absence of a particular base. Subsequently, the encoded features were reintegrated into the feature table, and the original categorical features were removed. Table 4.1 and Table 4.2 show the summary of the features before and after encoding in both human and cattle. The last column in Table 4.1 indicates the names of the various feature sets used in different experiments.

**Table 4. 1 Summary of human features before and after encoding**

| | Annotation | Number of features | Encoding approach | Number of columns after encoding | Feature set name |
|---|---|---|---|---|---|
| Sequence conservation | phastCons* | 2 | - | 2 | General features |
| | phyloP* | 2 | - | 2 | |
| Variant position properties | Distance to CpG island* | 1 | - | 1 | |
| | Distance to chromatin data | 1554 | - | 1554 | |
| | Distance to TSS* | 14 | - | 14 | |
| | Distance to regulatory features | 4 | - | 4 | |
| | Chromosome* | 1 | One-hot | 22 | |
| | Variant position* | 1 | - | 1 | |
| | Gene density (per mega base)* | 1 | - | 1 | |
| VEP annotation | Consequence* | 1 | One-hot | 51 | |
| Sequence context | Allele change* | 1 | Base-mapping | 8 | |
| | 5-mer flanking sequence* | 1 | Base-mapping | 20 | |
| Enformer | Predicted functional genomic score | 5313 | - | 5313 | Enformer features |
| EpiMap | Epigenomic data | 593 | - | 593 | EpiMap features |

** Annotations available for both humans and cattle*

**Table 4. 2 Summary of cattle features before and after encoding**

| | Annotation | Number of features | Encoding approach | Number of columns after encoding |
|---|---|---|---|---|
| Sequence conservation | phastCons | 1 | - | 1 |
| | phyloP | 1 | - | 1 |
| Variant position properties | Distance to CpG island | 1 | - | 1 |
| | Distance to TSS | 3 | - | 3 |
| | Chromosome | 1 | One-hot | 29 |
| | Variant position | 1 | - | 1 |
| | Gene density (per mega base) | 1 | - | 1 |
| VEP annotation | Consequence | 1 | One-hot | 31 |
| Sequence context | Allele change | 1 | Base-mapping | 8 |
| | 5-mer flanking sequence | 1 | Base-mapping | 20 |
| Enformer | Predicted functional genomic score | 5313 | - | 5313 |

### 4.2.3.3 Feature selection

The Enformer feature set consisted of a high-dimensional feature table with 5,313 columns, leading to increased computational complexity. Furthermore, this feature set might encompass redundant features that could potentially impact model performance. To address these challenges, a feature selection process was undertaken. The Enformer feature set first underwent the VarianceThreshold method in sklearn (Pedregosa et al., 2011) with a default threshold of 0 to remove all low-variance features. Then the sklearn mutual_info_classif package was employed to measure the association between each feature and the target variable (i.e., predicted as regulatory variant). Features exhibiting a mutual information value of 0 were excluded, as they demonstrated minimal correlation with the target variable.

After applying the variance and mutual information filters, a further feature selection process was conducted using the feature_importances_ attribute inherent in some machine learning algorithms. The sklearn SelectFromModel package was employed, with the tree-based machine learning algorithm CatBoost designated as the estimator. This

approach only retains those features whose feature_importances_ value derived from the model exceeds a predetermined threshold. To determine this threshold, a preliminary and an elaborate selection process were performed. First, the threshold range was defined spanning from 0 to the maximum feature_importances_ value. A learning curve was generated, plotting the threshold values on the x-axis against the cross-validation scores of the model on the y-axis, thereby pinpointing a narrower range for the optimal threshold. Subsequently, a similar but more elaborate process was performed, narrowing down the threshold range based on the outcomes of the preliminary process. Finally, a threshold value of 0.056 was picked as the feature_importances_ cut-off for Enformer features.

### 4.2.4 Machine learning models

#### 4.2.4.1 Machine learning model construction and training

Seven different machine learning algorithms, including Random Forest (Breiman, 2001), CatBoost (Prokhorenkova et al., 2017), XGBoost (T. Chen & Guestrin, 2016b), LightGBM (Ke et al., 2017), Gradient Boosting Decision Trees (GBDT) (Friedman, 2001), and Support Vector Machines (SVM) (Cortes & Vapnik, 1995) were selected as potential candidates for constructing the models. Random Forest, SVM, and GBDT models were instantiated using packages provided by sklearn. CatBoost, XGBoost, and LightGBM were built using their corresponding Python packages. The entire variant set, including both foreground and background variants, was randomly divided into a training and test set at a ratio of 70:30 using the sklearn train_test_split package. The training set was for model training and the separate test set was reserved for evaluating the models' performance on new, unseen data instances. To train the models and assess their performance in a less biased manner, as well as to maximize the effective utilization of the entire training data, a 5-fold cross-validation approach was employed.

#### 4.2.4.2 Hyper-parameter tuning

Each machine learning algorithm has some hyper-parameters that influence the learning process and impact the final model parameters. Adjusting these hyper-parameters can enhance the model's performance. Initially, the hyper-parameter tuning of the CatBoost

model was conducted using the RandomizedSearchCV and BayesSearchCV approaches from the scikit-optimize v0.9.0 library (Head et al., 2018), a Python library for hyper-parameter optimization. The BayesSearchCV method, which performs Bayesian optimization over hyper-parameters, demonstrated a more efficient tuning process and was subsequently employed for hyper-parameter tuning of all other models. For each algorithm, the search space for the hyper-parameters was defined as a dictionary, and the BayesSearchCV function was invoked to explore the target space and determine the optimal set of hyper-parameters that maximize the model's performance.

#### 4.2.4.3 Model ensemble

Further enhancement of model performance can be achieved through ensemble methods that combine individual models. The VotingClassifier from sklearn was utilized to organize three machine learning models: CatBoost, XGBoost, and LightGBM, which demonstrated superior performance among the seven algorithms considered. The VotingClassifier employs a majority voting strategy to determine the final prediction based on the predictions generated by the base models. Two options within the VotingClassifier were explored: the hard voting strategy, which makes final decisions by aggregating the majority predicted class among the base models' predictions, and the soft voting strategy, which employs the prediction probabilities of the base models to make final decisions.

#### 4.2.4.4 Incremental learning

When attempting to leverage human data to aid cattle regulatory prediction, apart from using human annotations or a human model, the incremental learning strategy was also explored. Incremental machine learning refers to an approach that progressively incorporates new data into the model, with the goal of learning from the new data while retaining the knowledge acquired from the previous data (Wu et al., 2019). In this study, the human model, trained on the annotations that are available in both human and cattle, was saved and subsequently employed as the initial model for a second round of training using cattle annotations. The ultimate model was then assessed by predicting cattle regulatory variants.

### 4.2.5 Machine learning model performance evaluation and interpretation

To evaluate the model's performance, various metrics were employed. In addition to the model's accuracy, other metrics such as precision, recall, and F1-score were also calculated to assess the model's proficiency in true positive predictions. Additionally, the Receiver Operating Characteristic (ROC) curve was generated, and the Area Under this Curve (AUROC) was computed to evaluate the model's discriminative capacity between regulatory variants and other variants.

Furthermore, model interpretation is a crucial step in the application of machine learning to address genomics challenges. It not only enhances the credibility of the model's predictions but also aids in comprehending the underlying characteristics captured by models from extensive data. To facilitate model interpretation, the feature importance was extracted from the model. Additionally, the SHapley Additive exPlanations (SHAP v0.39.0) approach (Lundberg & Lee, 2017) was utilized to elucidate the relationship between feature values and model outputs.

### 4.2.6 GWAS catalog data for human model application

To explore the potential application of the human model, variants from the GWAS catalog were utilized to determine if certain reported variants associated with traits have predicted regulatory effects from the trained model. Since not all lead variants from GWAS are causal in these associations, datasets from Open Target Genetics (Ghoussaini et al., 2021) were utilized to expand the dataset. This expansion includes all tag variants within each associated locus, where tag variants refer to those defined in the fine-mapping credible sets, as opposed to only lead variants from the GWAS catalog. The datasets were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/opentargets/genetics/. Then the Python package pyarrow (Richardson et al., 2023) was utilized to convert the data in *parquet* format to Python data frame. Following this, the variants were restricted to +/- 1Mb around the TSS and annotated using the variant annotation pipeline. The human model trained using the high-confidence data was utilized for prediction. To make a comparison, an equivalent number of background variants was randomly sampled from

the remaining variants in the human 1000 genomes cohort (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) that do not belong to the foreground variant dataset. These background variants were also restricted to +/- 1Mb around the TSS as the foreground data.

### 4.2.7 Functional genomic data for predicted cattle regulatory variants validation

To validate the potential effects of the cattle regulatory variants predicted by the model, I utilized the PRO-Cap data generated by Dr Rachel Owen in our group. Our group conducted PRO-Cap in two cattle aortic endothelial cells. Adapter sequence trimming was performed using fastp, a tool designed for pre-processing of fastq files (S. Chen et al., 2018). Next, UMI-tools was utilized for barcode extraction. UMI-tools is developed for handling Unique Molecular Identifiers (UMI) or single cell RNA-Seq cell barcodes (Smith et al., 2017), and in this case random UMIs were added to each fragment so that duplicates could be identified and discarded. The barcode patterns for the reads in the UMI-tools command were set to NNNNNNX where N represents the random part and X denotes the fixed part. Following this, the Spliced Transcripts Alignment to a Reference (STAR) software (Dobin et al., 2013) was used to generate the cattle genome indices (ARC-UCD1.2) and map the pre-processed PRO-cap reads to the reference genome. The output bam files underwent the deduplication procedure using the dedup tool in the UMI-tools and corresponding index files generated using samtools (Danecek et al., 2021). Subsequently, bamCoverage (Ramírez et al., 2016b) was employed to generate the BigWig format track. Finally, the peak caller PINTS was utilized to call the peaks for the PRO-Cap data. To obtain the predicted regulatory variants that intersected with the PRO-Cap peaks, the called peaks and the predicted variants were converted into GRanges objects in R and the subsetByOverlaps function in the GenomicRanges (Lawrence et al., 2013) R package was employed.

To further support the PRO-Cap results, TSS regions in cattle heart tissue defined using Cap Analysis Gene Expression sequencing (CAGE) were downloaded from https://api.faang.org/files/trackhubs/BOVREG_CAGE_EUROFAANG/ARS-

UCD1.2/tissues_TSS/ (Salavati et al., 2023). Since the paper did not provide data for aorta tissue, the most relevant tissue, heart, was used for this analysis.

## 4.3 Results

This section began with an exploration of the underlying genomic characteristics associated with regulatory variants. Subsequently, the prediction results for humans were presented and analysed, followed by the work conducted in predicting cattle regulatory variants. Figure 4.2 shows the different feature sets and models used in human and cattle predictions. For further clarity, a more detailed summary of the machine learning algorithms, variant sources, and features utilized in different experiments can be found in Table S4.1 and Table S4.2, respectively.

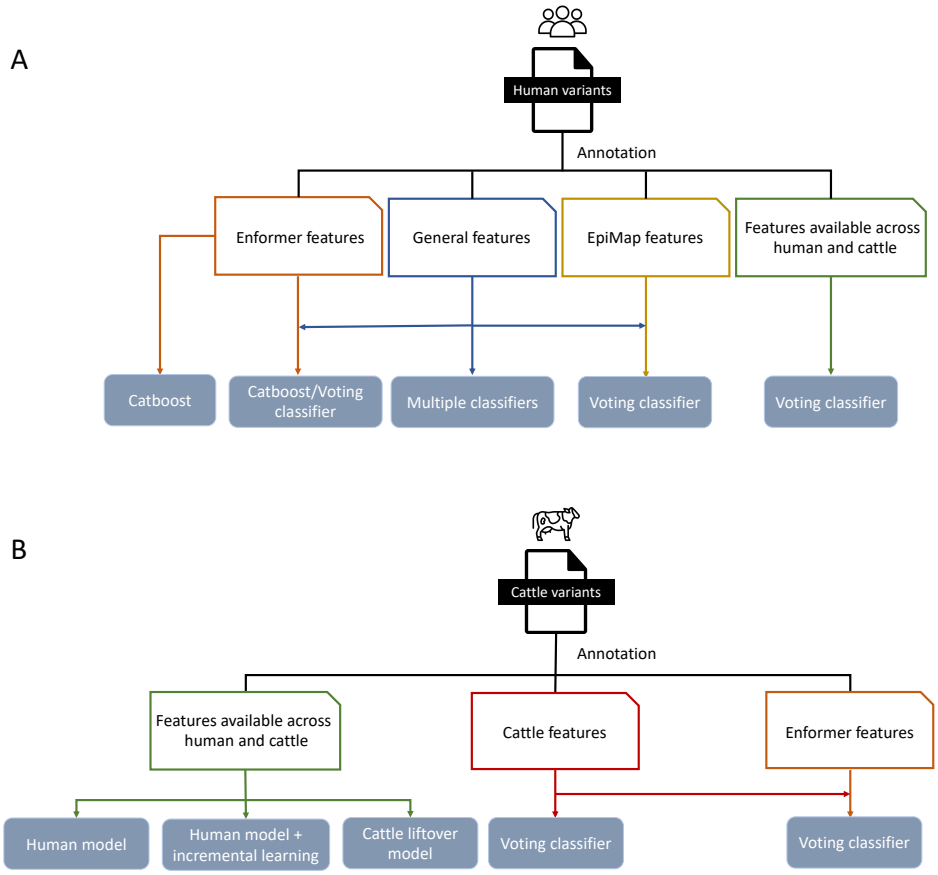



**Figure 4.2 Different feature sets and models used in the experiments in (A) human and (B) cattle.** (A) The multiple classifiers include Random Forest, Catboost, XGboost, LightGBM, GBDT, SVM. The green box in (B) represents three cattle regulatory variant prediction strategies based on human data. These strategies involve: utilizing pre-trained human models for cattle regulatory

variant prediction, employing incremental learning based on the human model for cattle prediction, and lifting over cattle variants to the human genome to train the model using human annotations. The red box in (B) represents the strategy based on cattle annotations directly.

### 4.3.1 Variant annotations analysis

#### 4.3.1.1 Cattle conservation scores based on Cactus alignment

Despite the potential importance of conservation scores at marking functional variants in the genome, phastCons and phyloP, two of the most widely used metrics, are not currently available for cattle. Therefore, I used my workflow from Chapter 2 to first generate and assess phastCons and phyloP scores generated from a cactus alignment of 241 mammals. As shown in Figure 4.3 (A), consistent with these metrics marking functional regions, coding regions exhibited a higher likelihood of being enriched near a phastCons score of 1, while non-coding regions were more likely to be enriched near a phastCons score of 0. This observation is consistent with the fact that coding regions generally display greater conservation compared to non-coding regions within the genome. Furthermore, this was reinforced by the distribution difference of the phyloP scores between coding and non-coding regions. These results collectively affirmed the dependability of my novel cattle conservation scores based on a large Cactus alignment and potential utility for discriminating between functional and background regions.

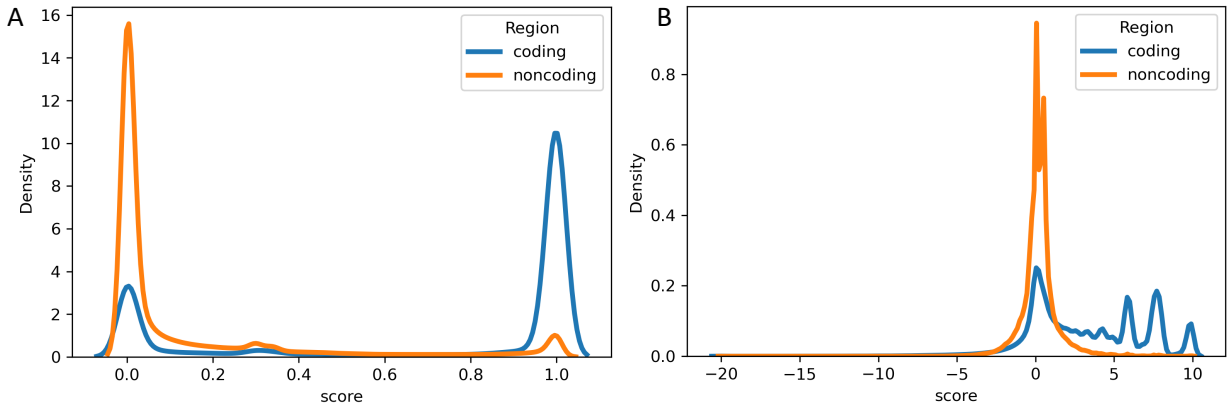



**Figure 4.3 Conservation score distribution differences in coding and non-coding regions in the cattle genome.**

(A) Kernel density estimate (KDE) plots of phastCons241way conservation scores in coding and non-coding regions. (B) KDE plots of phyloP241way conservation scores in coding and non-coding regions.

#### 4.3.1.2 Human and cattle annotations distribution analysis

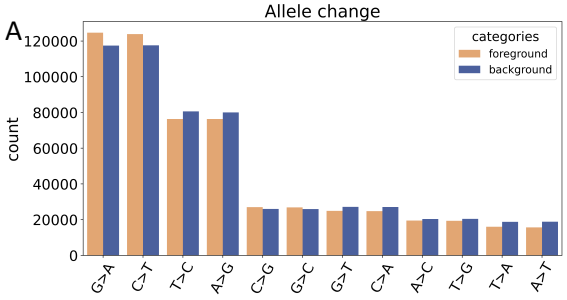
The distribution of annotations in both foreground and background datasets were investigated to determine if they could provide potential insights into the underlying characteristics of regulatory variants in humans and cattle.

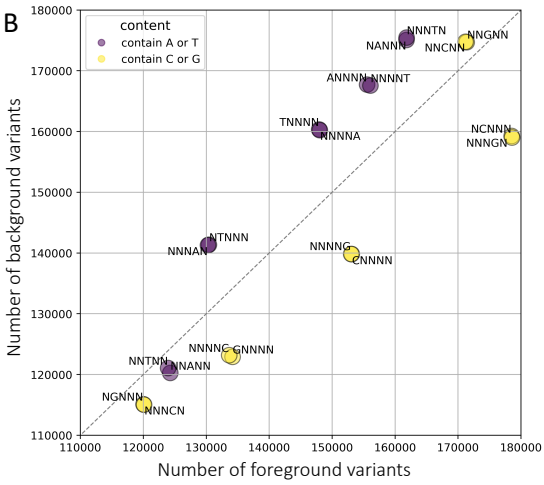
Figure 4.4 depicts the differences in the distribution of various annotations between the foreground and background groups in humans, shedding light on several factors associated with regulatory variants. One characteristic of regulatory variants was their tendency to exhibit a specific sequence context. As shown in Figure 4.4 (A), regulatory variants displayed a higher occurrence of G to A and C to T allele changes. Figure 4.4 (B) illustrates that variations with A:T base pairs at the variant position (the central base in the 5-mer flanking sequence) were more likely to be regulatory variants. Moreover, apart from the variant position, other positions within the 5-mer flanking sequences of regulatory variants tended to contain C:G base pairs. Furthermore, regulatory variants were more prevalent in gene-dense regions and exhibited enrichment near specific genomic elements, including CpG islands, regulatory elements, and chromatin data (Figure 4.4 D-H).

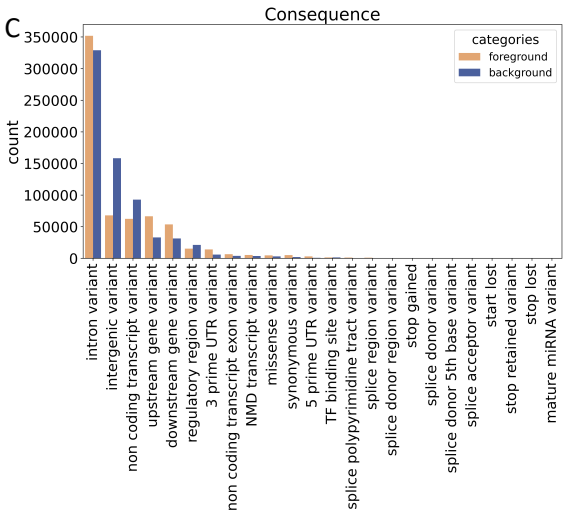
Upon comparing the distribution differences of conservation scores between foreground and background variants, a slight enrichment of regulatory variants was observed in regions exhibiting lower constraint (Figure 4.4 I, J). This finding is broadly consistent with the results reported by Mostafavi et al. in their study, where they highlighted that genes near eQTLs display less constraint (Mostafavi et al., 2022). This observation is perhaps counter-intuitive but potentially reflects the fact that genes near eQTLs possess greater flexibility in changing their expression.

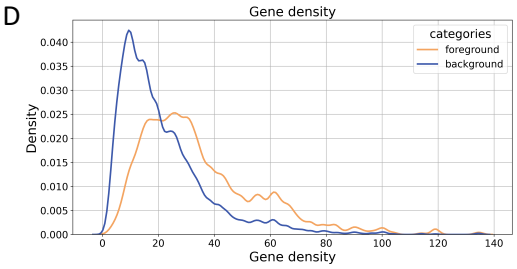
Figure 4.5 illustrates the characteristics of cattle regulatory variants. Most of the plots reveal similar distribution differences between regulatory and non-regulatory variants as in humans, highlighting shared characteristics of regulatory variants across different species. One noteworthy disparity between humans and cattle lies in the distribution of conservation scores. Specifically, cattle regulatory variants displayed a higher tendency to be enriched near a phastCons241way score of 1 (indicating higher conservation), while in human they were more likely to be enriched near a phastCons30way score of 0 (indicating lower conservation). This difference could be attributed to the varying numbers, categories, and evolutionary time scales of the species included in the multiple alignments used to calculate these conservation scores. The conservation scores for humans were computed based on the multiple alignment including 30 species, while the cattle conservation scores were derived from the 241-species multiple alignment. Another difference between humans and cattle was observed in their flanking sequences. Regulatory variants in humans tended to exhibit a higher proportion of C:G pairs compared to A:T at the variant position, whereas the trend is reversed in cattle.

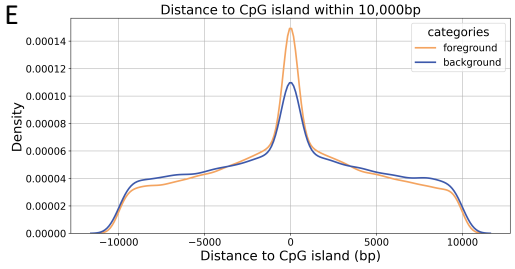
Despite the differences between human and cattle features, both species demonstrated notable distribution differences between foreground and background variants. These disparities provide a valuable opportunity for downstream machine learning models to discern patterns within different groups and consequently make new predictions.

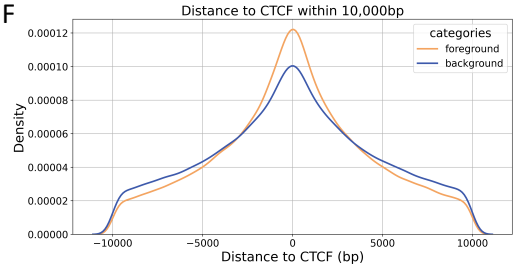

A
Allele change
count
categories
foreground
background
G>A
C>T
T>C
A>G
C>G
G>C
G>T
C>A
A>C
T>G
T>A
A>T

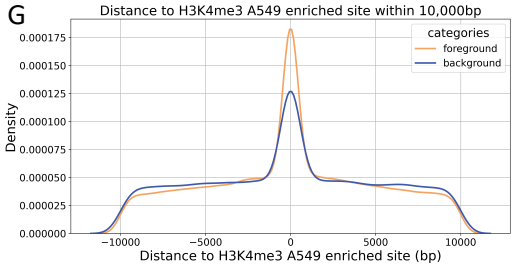

B
content
contain A or T
contain C or G
Number of background variants
Number of foreground variants
NNNTN
NANNN
NNCNN
NNGNN
ANNNN
NNNNT
TNNNN
NNNNA
NCNNN
NNNGN
NTNNN
NNNAN
NNNNG
CNNNN
NNNNC
GNNNN
NNTNN
NNANN
NGNNN
NNNCN

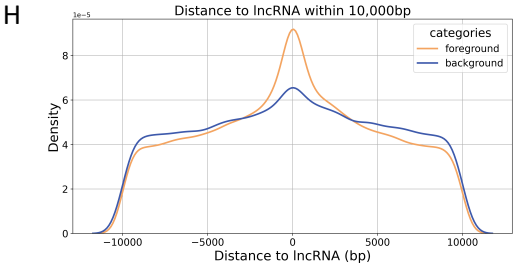

C
Consequence
count
categories
foreground
background
intron variant
intergenic variant
non coding transcript variant
upstream gene variant
downstream gene variant
regulatory region variant
3 prime UTR variant
non coding transcript exon variant
NMD transcript variant
missense variant
synonymous variant
5 prime UTR variant
TF binding site variant
splice polypyrimidine tract variant
splice region variant
splice donor region variant
stop gained
splice donor variant
splice donor 5th base variant
splice acceptor variant
start lost
stop retained variant
stop lost
mature miRNA variant

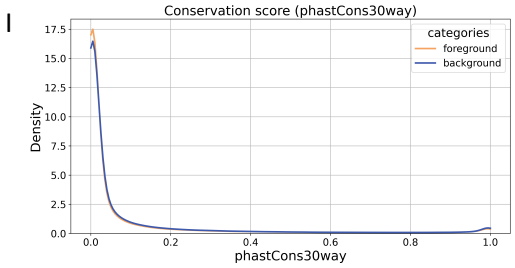

D
Gene density
Density
categories
foreground
background
Gene density


E
Distance to CpG island within 10,000bp
Density
categories
foreground
background
Distance to CpG island (bp)


F
Distance to CTCF within 10,000bp
Density
categories
foreground
background
Distance to CTCF (bp)


G
Distance to H3K4me3 A549 enriched site within 10,000bp
Density
categories
foreground
background
Distance to H3K4me3 A549 enriched site (bp)


H
Distance to lncRNA within 10,000bp
Density
categories
foreground
background
Distance to lncRNA (bp)


I
Conservation score (phastCons30way)
Density
categories
foreground
background
phastCons30way

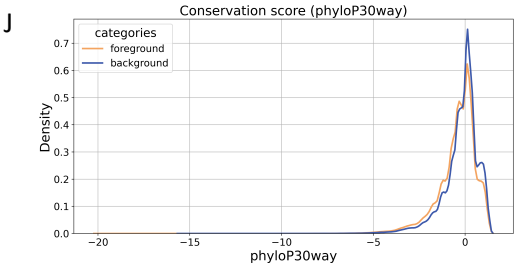

J
Conservation score (phyloP30way)
Density
categories
foreground
background
phyloP30way

**Figure 4.4 Characteristics of human regulatory variants.** (A) Number of foreground regulatory variants and background variants by their allele changes (reference > alternative). (B) Number of variants with different 5-mer flanking sequences between the foreground and background data. Each circle represents a type of 5-mer flanking sequence with a specific base at a particular position, and the colour of the circle indicates the sequence content. All types of 5-mer flanking sequences exhibit significant differences between the groups (Chi-Squared test, p-value < $\mathbf{1.5 \times 10^{-8}}$). (C) Number of foreground regulatory variants and background variants by their variant consequences. (D) Kernel density estimate (KDE) plots of gene density between foreground and background variants. Gene density displays significant differences between the groups (Two-sample Kolmogorov-Smirnov test, p-value < $\mathbf{5.87 \times 10^{-114}}$). (E-H) KDE plots of distances from foreground and background variants to different genomic elements within 10kb. These annotations show significant differences between the groups. (Two-sample Kolmogorov-Smirnov test, p-values < $\mathbf{3.34 \times 10^{-36}}$, $\mathbf{1.02 \times 10^{-8}}$, $\mathbf{1.59 \times 10^{-34}}$, $\mathbf{1.08 \times 10^{-8}}$, for panels E, F, G, and H respectively). (I-J) KDE plots of conservation scores for foreground and background variants. The phastCons30way and phyloP30way scores exhibit significant differences between the groups (Two-sample Kolmogorov-Smirnov test, p-values < $\mathbf{8.66 \times 10^{-6}}$ and $\mathbf{3.85 \times 10^{-26}}$, respectively).

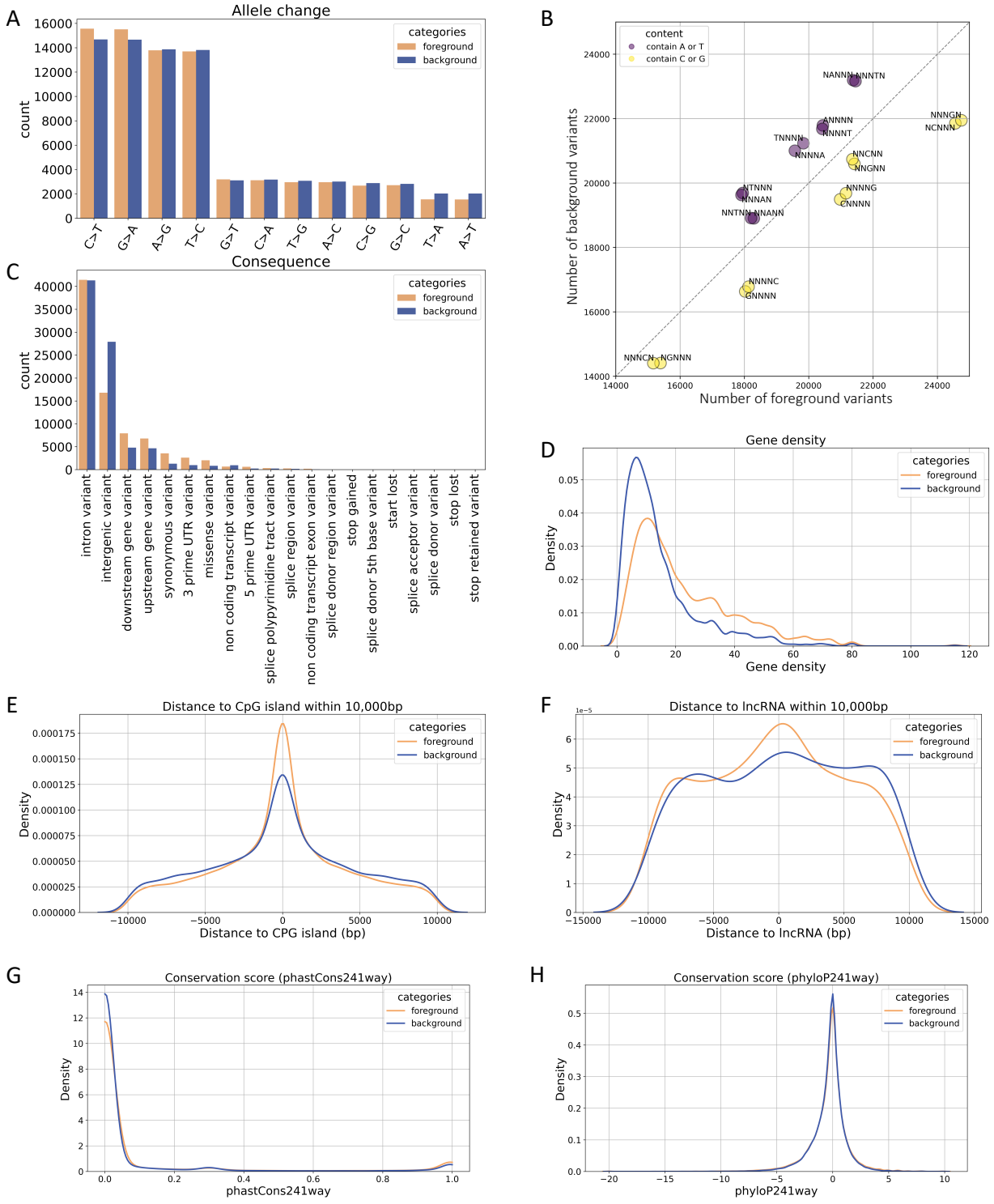



**Figure 4.5 Characteristics of cattle regulatory variants.** The variants were annotated using cattle annotations (A) Number of foreground regulatory variants and background variants by their allele changes (reference > alternative). (B) Number of variants with different 5-mer flanking sequences between foreground and background data. All types of 5-mer flanking sequences

exhibit significant differences between the groups (Chi-Squared test, p-value < $\mathbf{3.6 \times 10^{-8}}$). (C) Number of foreground regulatory variants and background variants by their variant consequences. (D) KDE plot of gene density between foreground and background variants. Gene density displays significant differences between the groups (Two-sample Kolmogorov-Smirnov test, p-value < $\mathbf{7.14 \times 10^{-81}}$). (E) KDE plot of distances from foreground and background variants to CpG islands within 10kb (Two-sample Kolmogorov-Smirnov test, p-value < $\mathbf{1.10 \times 10^{-16}}$). (F) KDE plot of distances from foreground and background variants to lncRNA (Two-sample Kolmogorov-Smirnov test, p-value < $\mathbf{2.40 \times 10^{-5}}$). (G) KDE plot for phastCons241way scores in foreground and background variants. (Two-sample Kolmogorov-Smirnov test, p-value < $\mathbf{1.69 \times 10^{-4}}$) (H) KDE plot for phyloP241way scores in two groups (Two-sample Kolmogorov-Smirnov test, p-value < $\mathbf{1.58 \times 10^{-3}}$).

### 4.3.2 Human regulatory variant prediction results

#### 4.3.2.1 Prediction results for regulatory variants across different tissues

To provide a baseline for my cattle models, I first developed machine learning models for predicting human regulatory variants. Given the available high quality training data for humans, this could provide a potential upper bound of the prediction accuracies I could likely obtain from other species such as cattle. Initially, the model training in humans did not take into account the tissue where the fine-mapped regulatory variant had been identified. In addition to the entire regulatory variant dataset, a high-confidence regulatory variant dataset was obtained by filtering the data according to their CaVEMaN causal probability with a threshold of 0.5, aiming to assess whether the model performs better with more reliable data. After filtering by causal probability, 45,987 variants were retained in the foreground dataset and the background dataset was randomly down-sampled to achieve an equivalent size. Finally, the size of the high-confidence feature table was 91,974 variants × 1,680 features (using general features as specified in Table 4.1), and the size of the entire feature table was 1,170,588 variants ×1680 features, where each row represents a variant, and each column represents an encoded feature.

The datasets were split into training and test sets at a ratio of 70% and 30%, respectively. A Random Forest was employed as the baseline model. Figure 4.6 displays the ROC

curves for the models trained on both the entire regulatory variant dataset and the high-confidence regulatory variant dataset, respectively. As expected, the model trained on the high-confidence data outperformed the one trained on the entire dataset, despite the latter containing approximately 12 times more training data, highlighting the importance of the training data quality in determining the final model performance.

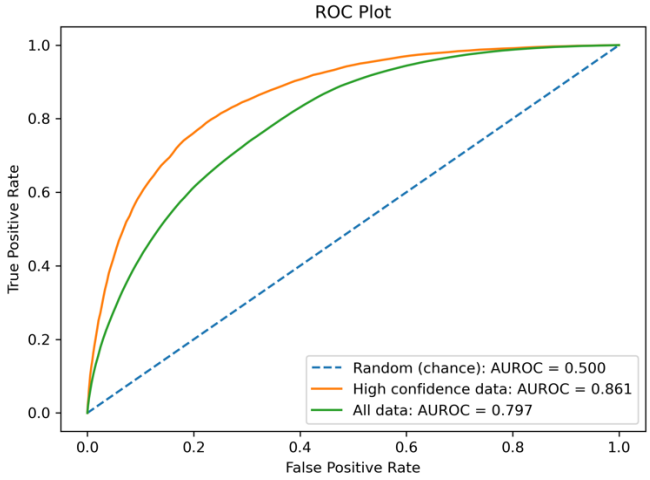



**Figure 4.6 Receiver Operating Characteristic (ROC) curves and corresponding Area Under the Receiver Operating Characteristic (AUROC) scores of the Random Forest model trained using the entire dataset and the high-confidence dataset.** The AUROC scores reflect the model's ability to distinguish between regulatory variants and background variants, with scores closer to 1 indicating better performance. The dashed blue line represents the expected performance from random guesses.

To determine which machine learning algorithms to employ in the study, an initial exploration was conducted, comparing the performance of several popular machine learning algorithms in predicting human regulatory variants. Seven different algorithms, including Random Forest, CatBoost, XGBoost, LightGBM, Gradient Boosting Decision Trees (GBDT), and Support Vector Machine (SVM), were trained on the high-confidence training set (64,381 variants) and tested on a separate test set (27,592 variants). All models were trained using their default parameters without hyper-parameter tuning. As depicted in Figure 4.7, the six tree-based machine learning models demonstrated consistently better performance with higher AUROC scores when compared to SVM. Taking into account both the model performance and interpretability, three tree-based

machine learning models, CatBoost, XGBoost, and LightGBM were selected for subsequent studies to perform various tasks.

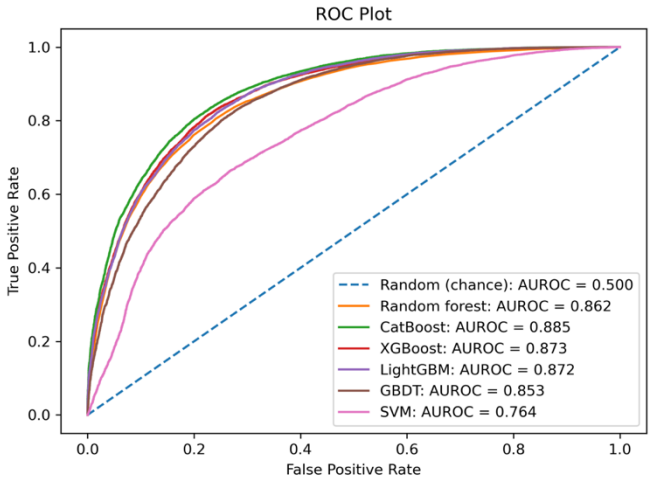



**Figure 4.7 ROC curves and AUROC scores for seven machine learning models trained and tested using the high-confidence data.**

To further improve the performance of the tree-based models, hyper-parameter tuning was employed to the models. And then the voting classifier was used to organize CatBoost, XGBoost, and LightGBM for further model ensemble. Table 4.3 summarizes the metrics for different models when training and testing on high confidence regulatory variants. The best-performing model was CatBoost, achieving an AUROC score of 0.898 after tuning. Additionally, the voting classifier, employing a soft strategy as introduced in Section 4.2.4.3, exhibited a slight improvement over the one using a hard strategy and other individual models. Therefore, the voting classifier using a soft strategy was employed as the model in the downstream analyses if not specified.

**Table 4.3 Different metrics for the models after tuning**

| Model | Recall | Precision | Accuracy | AUROC | Average cv score [a] | Standard deviation of cv scores [b] |
|---|---|---|---|---|---|---|
| CatBoost | 0.825 | 0.787 | 0.801 | 0.898 | 0.804 | 0.004 |
| XGBoost | 0.813 | 0.782 | 0.794 | 0.880 | 0.791 | 0.002 |
| LightGBM | 0.812 | 0.770 | 0.786 | 0.880 | 0.792 | 0.002 |
| Voting (soft) | 0.830 | 0.789 | 0.809 | 0.903 | 0.807 | 0.003 |
| Voting (hard) | 0.823 | 0.782 | 0.798 | - | 0.801 | 0.003 |

[a] *The average F1 score of the models obtained through 5-fold cross-validation.*
[b] *The standard deviation of the scores obtained from 5-fold cross-validation for the models*

To gain a deeper understanding of how features influence prediction outcomes, a SHAP summary plot was generated for model interpretability. Figure 4.8 presents the top 30 most important features for predicting high confidence regulatory variants, along with their respective impacts on the model's prediction. The foremost influential feature is the distance to the promoter. As depicted in the plot, a moderate distance-to-promoter value (around 0, as distance includes both positive and negative values), i.e., proximity to the promoter region, is associated with a higher probability of a variant being predicted as a regulatory variant. Conversely, a high or low distance-to-promoter value, indicating greater distance either upstream or downstream from the promoter, is linked to a higher probability of being predicted as a non-regulatory variant.

The second most prominent feature is gene density, where a higher gene density corresponds to an elevated SHAP value. This indicates that increased gene density is linked to an increased likelihood of a variant being predicted as a regulatory variant. This aligns with the observation that regulatory variants tend to be more prevalent in gene-dense regions. Moreover, the categorical feature consequence (intergenic variant) emerges as the third most important feature. As described in the feature encoding section, categorical features were encoded into strings and subsequently split into binary columns containing values of 0 or 1. For the consequence feature, each column post-encoding denotes a different category of variant consequence. From the SHAP plot, it is evident

that an intergenic variant column value of 0, meaning not an intergenic variant, is correlated with a higher probability of a variant being classified as a regulatory variant.

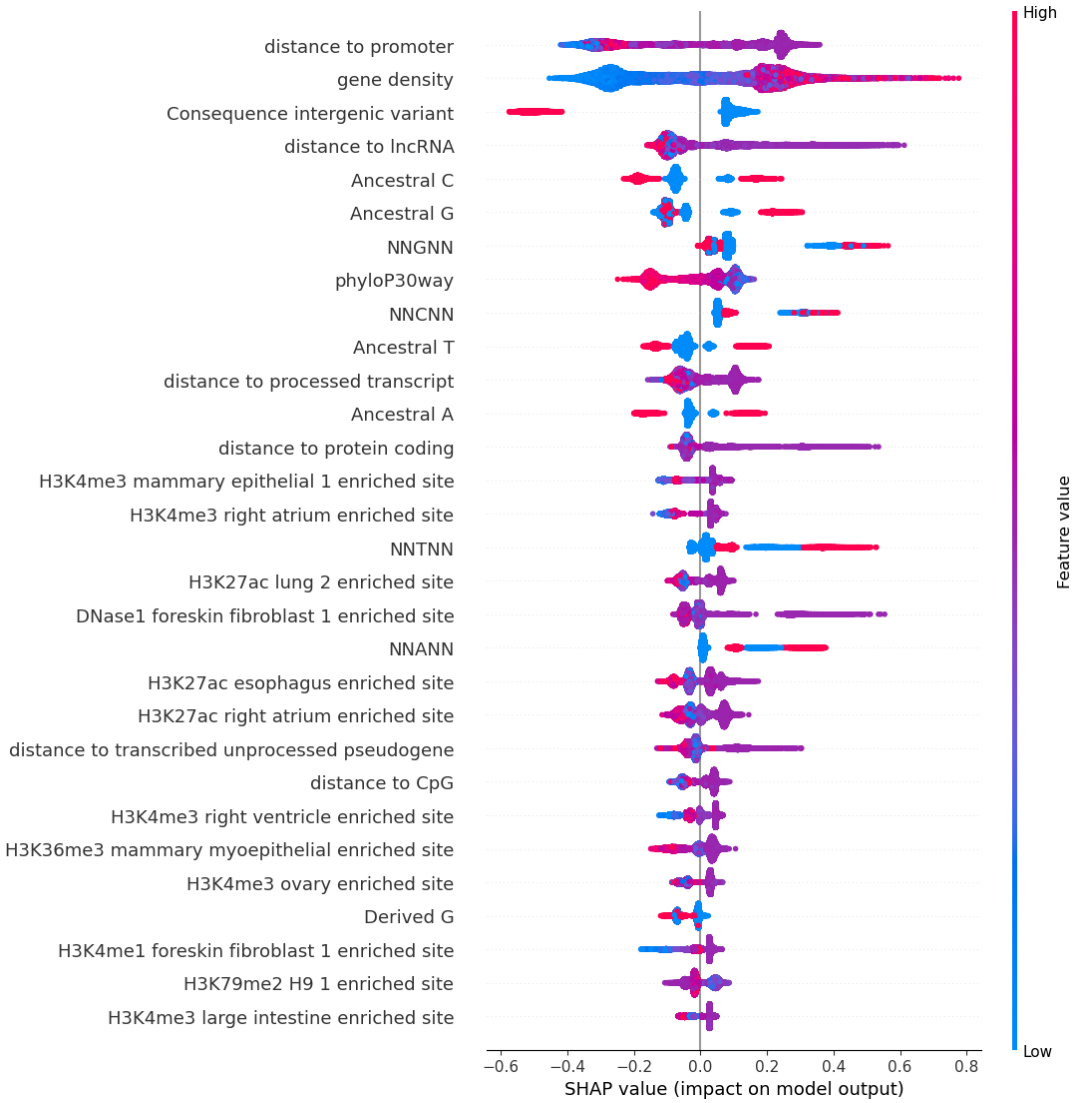



**Figure 4.8 SHAP plot of the model trained and tested using high confidence regulatory variants.** The top 30 features are ranked in descending order based on their feature importance in the model. The colour bar indicates high and low values of the features. The x axis shows the effect of the feature value on the model's prediction with a positive SHAP value associated with an increased probability of being called a regulatory variant.

### 4.3.2.2 Comparison of model performance based on different feature sets

Subsequently, I explored the utility of Enformer features in predicting regulatory variants. Since Enformer models can directly predict gene expression data from DNA sequences, it is feasible to obtain predicted gene expression data for species lacking these data using the Enformer model trained on human data. The exploration of utilizing Enformer features for predicting regulatory variants in humans is also an investigation into the potential of employing predicted gene expression data in other species for regulatory variant prediction.

The high confidence dataset from above was further annotated using the Enformer sub-workflow, resulting in an Enformer feature table with a size of 91,974 variants × 5,313 features. Due to the constraints on available cores and memory in the Eddie environment, only the CatBoost model with default parameters could be trained with such a high-dimensional feature table. Therefore, CatBoost with default parameters was employed in this part. The CatBoost model was trained using the Enformer features, the general features (1680 columns after encoding), and a combination of both feature sets (6,993 columns), respectively. As shown in the ROC plot for these three experiments (Figure 4.9), the model trained using general features outperformed the one trained with Enformer features. Additionally, the model trained using the combined features exhibit comparable performance to the model trained with general features only. This suggests that including Enformer predictions in the model lead to limited improvement.

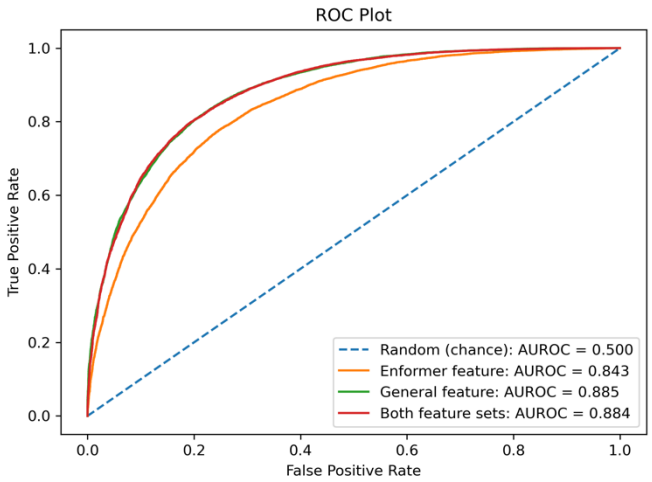


**Figure 4.9 ROC curves and AUROC scores for the CatBoost model trained using three different feature sets.**

To assess whether the high-dimensional Enformer features include irrelevant or redundant features that could potentially impede model performance, a feature selection process was employed to the Enformer feature set. Initially, the 5,313-columns feature set underwent the Variance Threshold and Mutual Information filters with both thresholds set to 0, resulting in the removal of 164 features from the original set. Subsequently, the remaining features were subjected to the SelectFromModel method for further selection, with CatBoost as the estimator. After filtering, 440 features were retained in the Enformer feature set. The CatBoost model was then trained on the 440 selected Enformer features and the combined feature set, which included the selected Enformer features along with the general features. The AUROC scores were 0.840 and 0.885 respectively, which were comparable to using the entire Enformer feature set. Subsequently, the previously mentioned voting classifier was trained using this combined feature set, resulting in an AUROC of 0.899. This performance closely paralleled the AUROC of the voting model trained on the general feature set (AUROC=0.903), as detailed in Table 4.3. These findings underscore that the Enformer features did not contribute additional discriminative information to the model's ability to predict regulatory variants, beyond what was already provided by the general features.

In addition to the Enformer predicted chromatin features, another set of directly assayed chromatin features from EpiMap was also explored in an attempt to improve the model's performance on predicting high-confidence data. The EpiMap feature set consisted of 593 columns. The model was trained using a combination of EpiMap features and the general features, resulting in an AUROC score of 0.898, which was also comparable with the model trained using only the general features.

The comparable performance of the models trained using these different feature sets suggests that the general features employed in this study already encompass abundant information that is conducive for regulatory variant prediction. The addition of

complementary annotations from the popular deep-learning-based architecture Enformer to the comprehensive high-resolution epigenomic annotations of EpiMap may offer limited benefits to the model.

#### 4.3.2.3 Prediction results for tissue-specific data

One potential reason that the assayed or inferred chromatin data was not improving the model was due to fitting a tissue-agnostic model i.e., regulatory variants from across tissues were combined into a single set. As regulatory variants and chromatin data can be tissue-specific, this may be affecting model performance. In this section, I conducted training and testing of the models across different tissues to assess their performance. Fine-mapped variants from five different tissues of relevance to cattle research, namely blood, liver, adipose, muscle and breast, were extracted from the CaVEMaN dataset. Given the potential scarcity of training data, the tissue-specific data were not filtered according to the CaVEMaN causal probability. Then the background variants were down-sampled to match the number of foreground variants retained in each tissue. Consequently, the adipose, liver  blood, breast, and muscle sets included 64,378, 24,416, 48,774, 42,022, 54,244 variants, respectively. Each set was split into training and test sets at a ratio of 70% and 30%. Then the voting classifier was trained on the training set and tested on the test set. As depicted in Table 4.4, the liver-based model outperformed the other tissue-specific models, achieving an AUROC of 0.919. This performance was also superior to the model based on the high-confidence cross-tissue data with an AUROC score of 0.903. Furthermore, these human tissue-specific models were also trained using just the annotations that were available across both human and cattle (as specified in Table 4.1) for further comparison with later cattle tissue-specific models (Table 4.5). Again, the liver model had the highest accuracy.

**Table 4.4 Different metrics for human tissue-specific models**

| Tissue | Recall | Precision | Accuracy | AUROC | Average cv score [a] | Standard deviation of cv scores [b] |
|---|---|---|---|---|---|---|
| Blood | 0.841 | 0.778 | 0.801 | 0.890 | 0.808 | 0.003 |
| Liver | 0.853 | 0.812 | 0.829 | 0.919 | 0.832 | 0.002 |
| Adipose | 0.823 | 0.774 | 0.790 | 0.874 | 0.798 | 0.003 |
| Muscle | 0.823 | 0.769 | 0.788 | 0.875 | 0.795 | 0.003 |
| Breast | 0.846 | 0.800 | 0.816 | 0.901 | 0.822 | 0.002 |

**Table 4.5 Different metrics for human tissue-specific models using cattle-available annotations**

| Tissue | Recall | Precision | Accuracy | AUROC | Average cv score [a] | Standard deviation of cv scores [b] |
|---|---|---|---|---|---|---|
| Blood | 0.829 | 0.764 | 0.787 | 0.868 | 0.795 | 0.004 |
| Liver | 0.838 | 0.793 | 0.811 | 0.896 | 0.815 | 0.002 |
| Adipose | 0.802 | 0.751 | 0.767 | 0.849 | 0.776 | 0.003 |
| Muscle | 0.806 | 0.756 | 0.773 | 0.853 | 0.781 | 0.002 |
| Breast | 0.826 | 0.781 | 0.795 | 0.877 | 0.803 | 0.003 |

[a] *The average F1 score of the models obtained through 5-fold cross-validation.*
[b] *The standard deviation of the scores obtained from 5-fold cross-validation for the models*

Subsequently, I compared the models' performance on variants obtained from five distance-to-TSS ranges across different tissues. These TSS bins were selected to ensure that each bin contains roughly equal number of variants in each tissue. Both the tissue-specific models and the high-confidence model exhibited decreased performance when using the variants that were farther away from the TSS (Figure 4.10). This suggests that the models are more effective at capturing relationships and making accurate predictions for variants that are closer to the TSS. Moreover, among these various models, the liver model consistently demonstrated good overall performance across different ranges, particularly when dealing with variants that are located farther away from the TSS. In other words, the liver model performed well in predicting regulatory variants at distal regulatory elements, such as enhancers. This may be attributed to the liver being a

relatively more homogeneous tissue compared to the other tissue, where the other tissues comprise a mixture of signals from various tissues. As a result, the model might find it easier to capture enhancer signals within the liver.

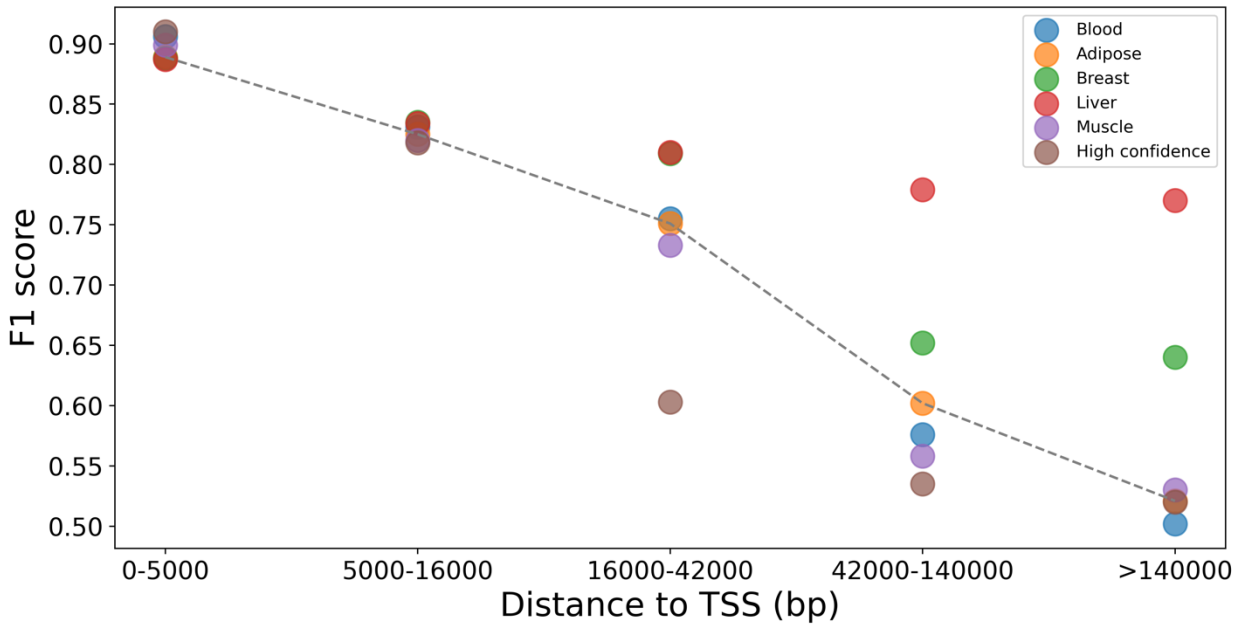
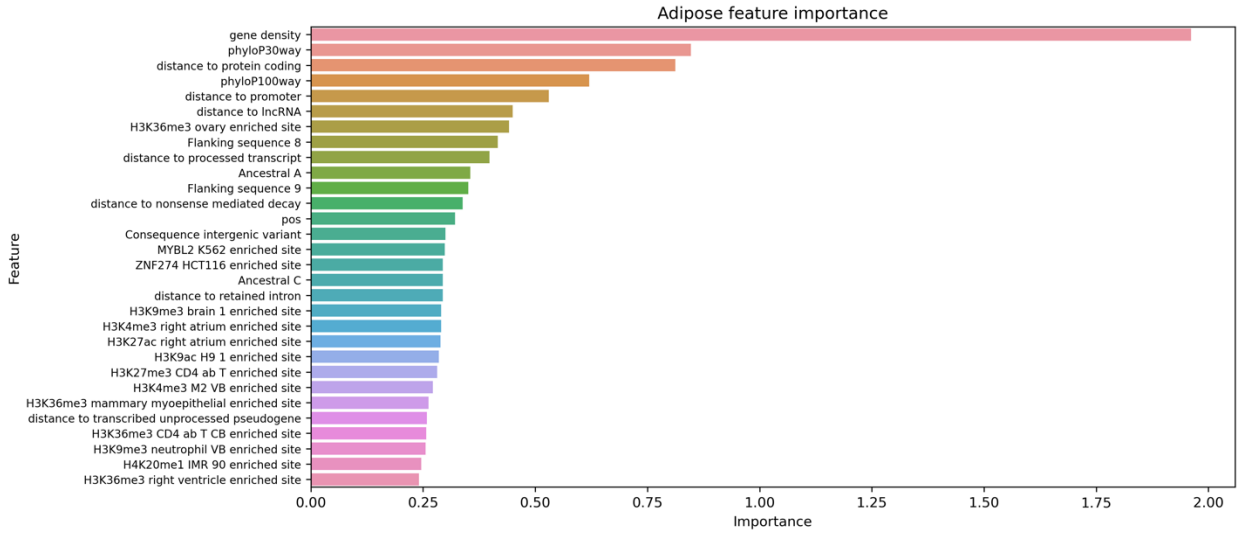


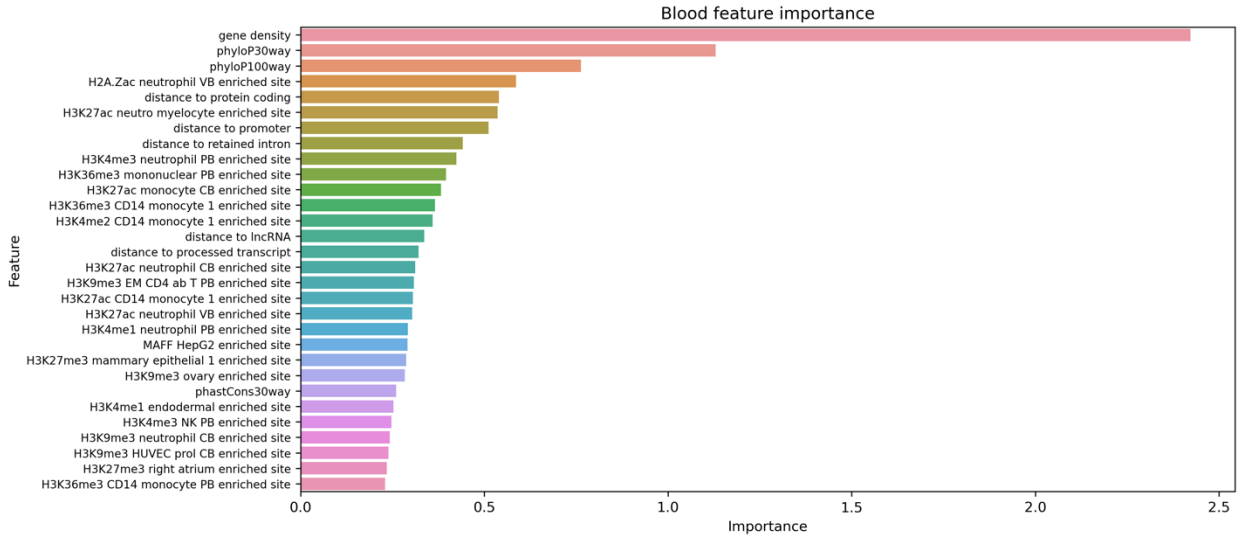

**Figure 4.10 Performance of models when training and testing using variants within different distance ranges to the TSS.** The grey dashed line is plotted based on the median F1 score among the six datasets in each range.

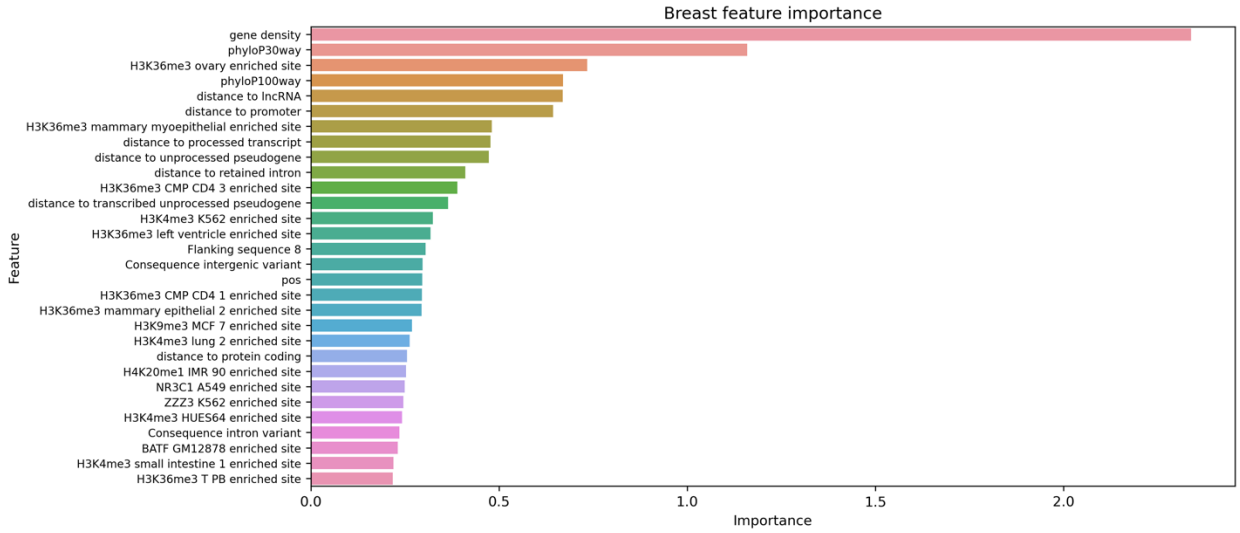
The feature importance of each tissue-specific model was obtained, and Figure 4.11 shows the top 30 most important features in each tissue-specific model. In addition to features that exhibit importance across different models, such as gene density and distance to TSS, certain features emerged solely among the top features in specific tissue models. For example, the distance from the variant to open chromatin regions marked by H2A.Zac in the neutrophil cell line ranked among the top 5 most important features exclusively in the blood-specific model. Similarly, the distance from variant to open chromatin regions marked by H3K4me1 in the psoas muscle featured among the top 10 most important attributes in the muscle-specific model. This suggests that tissue-specific chromatin data can improve tissue-specific model predictions, but at the same time the

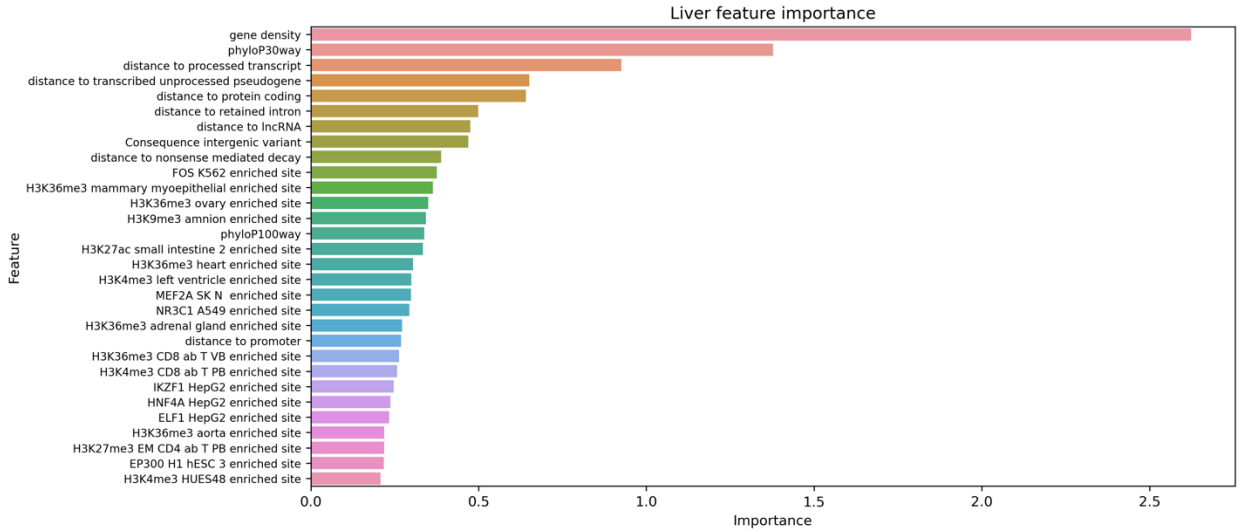
improvement over the general model not including such features is relatively small (Table 4.4 vs 4.5).

The improved model performance in tissue-specific models, in conjunction with the presence of distinct top features specific to certain tissues, highlights the importance of considering tissue context for accurately predicting regulatory variants.




Breast feature importance
gene density
phyloP30way
H3K36me3 ovary enriched site
phyloP100way
distance to lncRNA
distance to promoter
H3K36me3 mammary myoepithelial enriched site
distance to processed transcript
distance to unprocessed pseudogene
distance to retained intron
H3K36me3 CMP CD4 3 enriched site
distance to transcribed unprocessed pseudogene
H3K4me3 K562 enriched site
H3K36me3 left ventricle enriched site
Flanking sequence 8
Consequence intergenic variant
pos
H3K36me3 CMP CD4 1 enriched site
H3K36me3 mammary epithelial 2 enriched site
H3K9me3 MCF 7 enriched site
H3K4me3 lung 2 enriched site
distance to protein coding
H4K20me1 IMR 90 enriched site
NR3C1 A549 enriched site
ZZZ3 K562 enriched site
H3K4me3 HUES64 enriched site
Consequence intron variant
BATF GM12878 enriched site
H3K4me3 small intestine 1 enriched site
H3K36me3 T PB enriched site
Feature
0.0
0.5
1.0
1.5
2.0
Importance


Liver feature importance
gene density
phyloP30way
distance to processed transcript
distance to transcribed unprocessed pseudogene
distance to protein coding
distance to retained intron
distance to lncRNA
Consequence intergenic variant
distance to nonsense mediated decay
FOS K562 enriched site
H3K36me3 mammary myoepithelial enriched site
H3K36me3 ovary enriched site
H3K9me3 amnion enriched site
phyloP100way
H3K27ac small intestine 2 enriched site
H3K36me3 heart enriched site
H3K4me3 left ventricle enriched site
MEF2A SK N enriched site
NR3C1 A549 enriched site
H3K36me3 adrenal gland enriched site
distance to promoter
H3K36me3 CD8 ab T VB enriched site
H3K4me3 CD8 ab T PB enriched site
IKZF1 HepG2 enriched site
HNF4A HepG2 enriched site
ELF1 HepG2 enriched site
H3K36me3 aorta enriched site
H3K27me3 EM CD4 ab T PB enriched site
EP300 H1 hESC 3 enriched site
H3K4me3 HUES48 enriched site
Feature
0.0
0.5
1.0
1.5
2.0
2.5
Importance

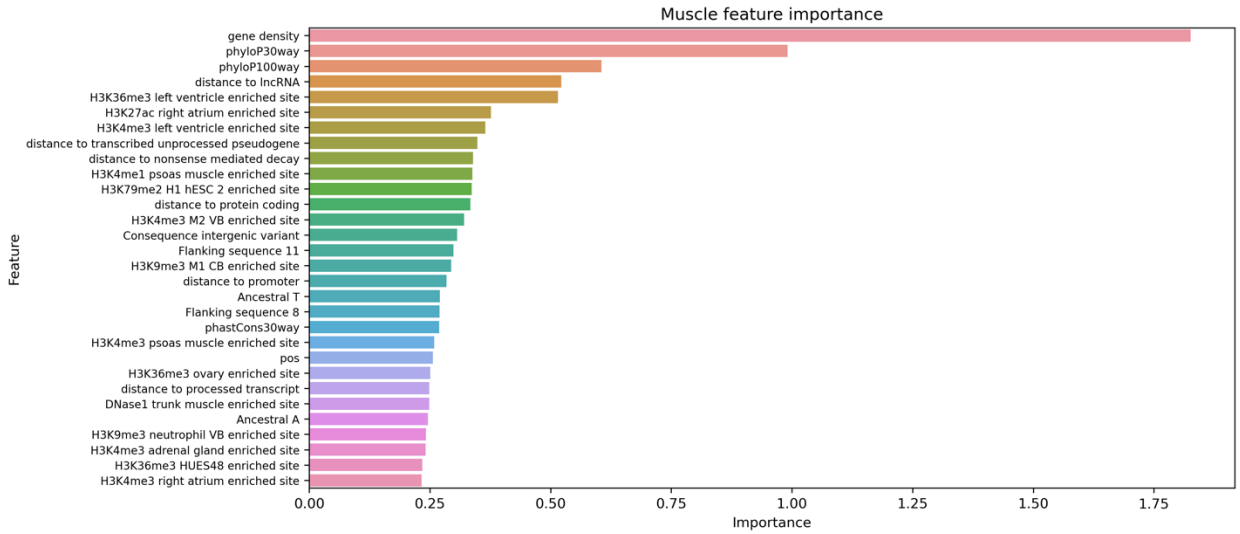

Muscle feature importance
gene density
phyloP30way
phyloP100way
distance to lncRNA
H3K36me3 left ventricle enriched site
H3K27ac right atrium enriched site
H3K4me3 left ventricle enriched site
distance to transcribed unprocessed pseudogene
distance to nonsense mediated decay
H3K4me1 psoas muscle enriched site
H3K79me2 H1 hESC 2 enriched site
distance to protein coding
H3K4me3 M2 VB enriched site
Consequence intergenic variant
Flanking sequence 11
H3K9me3 M1 CB enriched site
distance to promoter
Ancestral T
Flanking sequence 8
phastCons30way
H3K4me3 psoas muscle enriched site
pos
H3K36me3 ovary enriched site
distance to processed transcript
DNase1 trunk muscle enriched site
Ancestral A
H3K9me3 neutrophil VB enriched site
H3K4me3 adrenal gland enriched site
H3K36me3 HUES48 enriched site
H3K4me3 right atrium enriched site
Feature
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Importance

**Figure 4.11 The top 30 most important features in 5 tissue-specific models.** The features are ranked in descending order based on their feature importance in the model.

### 4.3.2.4 Examples of GWAS variants with predicted regulatory effect

The human model trained using high-confidence data was employed to predict whether a GWAS variant associated with a trait could function as a regulatory variant. Among the 976,076 variants from Open Target Genetics (foreground variants), which encompass both lead and tag variants in GWAS loci, approximately 35% of the variants were predicted as regulatory variants (with a prediction probability greater than 0.5). Among the equivalent number of variants in the background dataset that were randomly sampled from the 1000 Genomes cohorts, approximately 22% of the variants were predicted as regulatory variants. Figure 4.12 (A) illustrates the probability distribution difference between foreground and background variants. Compared to the background variants, GWAS variants are more likely to have a higher probability of being predicted as regulatory variants. It should be noted that the allele frequency was not taking into consideration when sampling the background variants from the 1000 Genomes cohorts to match with the GWAS variants, which typically focus on common variants with minor allele frequencies greater than 5%.

In theory, the greater the ratios of GWAS variants to background variants predicted as regulatory variants, the better. Therefore, to explore the impact of the probability threshold for distinguishing regulatory variants in the model, which was set to 0.5 by default, different threshold was utilized, and the corresponding ratio of the proportion of foreground variants predicted as regulatory to the proportion of background variants predicted as regulatory was calculated, i.e., $\frac{number\ of\ foreground\ variants\ predicted\ as\ regulatory/number\ of\ all\ foreground\ variants}{number\ of\ background\ variants\ predicted\ as\ regulatory/number\ of\ all\ backgroung\ variants}$. Figure 4.12 (B) shows different thresholds and their corresponding ratios. As the threshold increases, the ratio also increase, suggesting a potential for improving the prediction results by adjusting the threshold. However, noise becomes apparent when the threshold exceeds

0.85. This is attributed to the dramatically reduced numbers of variants predicted as regulatory in both the foreground and background datasets as the threshold exceeds 0.85.

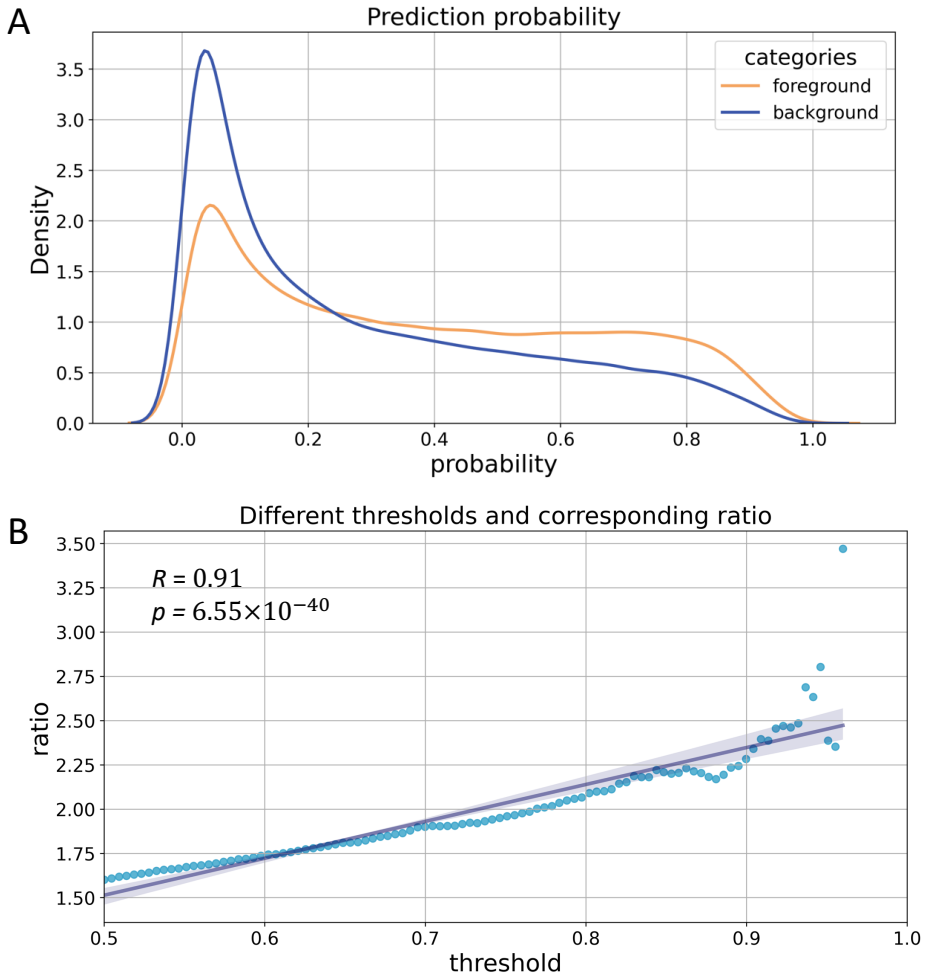



**Figure 4.12 Prediction results of GWAS and background variants.** (A) Prediction probability distribution of the GWAS variants (foreground variants) and background variants with a prediction probability threshold of 0.5. (B) Different thresholds and their corresponding ratios of the proportion of GWAS variants predicted as regulatory to the proportion of background variants predicted as regulatory. The line depicted on the plot represents the linear regression line fitted to the data points. The correlation coefficient and the corresponding p-value are specified in the plot.

Figure 4.13 shows the examples of three GWAS variants predicted as regulatory variants by the model, with prediction probabilities greater than 0.85. Variant rs3897478 (Figure

4.13 A) was located upstream of the *ADAM30* gene. This variant was identified as the lead variant in one of the loci associated with Crohn's disease, a form of inflammatory bowel disease (IBD) in humans (Jostins et al., 2012). The model predicted rs3897478 as a regulatory variant, suggesting its potential regulatory impact on the expression of the *ADAM30* gene, thereby contributing to the downstream phenotype.

Figure 4.13 (B) presents an example of a genomic locus containing a variant predicted as a regulatory variant with a prediction probability of 0.85. This genomic locus was found associated with smoking status measurement trait, which evaluates the genetic factors influencing tobacco use (M. Liu et al., 2019). Upon further exploration of variant rs13023601 in the Ensembl browser, it was discovered to overlap with an enhancer (ENSR00001037040). As a result, this variant may affect the function of this enhancer and consequently impact the expression of the *LINC01823* gene. Figure 4.13 (C) depicts another variant rs793191, associated with smoke initiation trait, predicted as a regulatory variant with a probability of 0.90 (Saunders et al., 2022). According to Ensembl, this variant was located in a promoter region (ENSR00000454531), suggesting its potential to influence the expression of the mapped gene *LINC02387* and contribute to the phenotype.

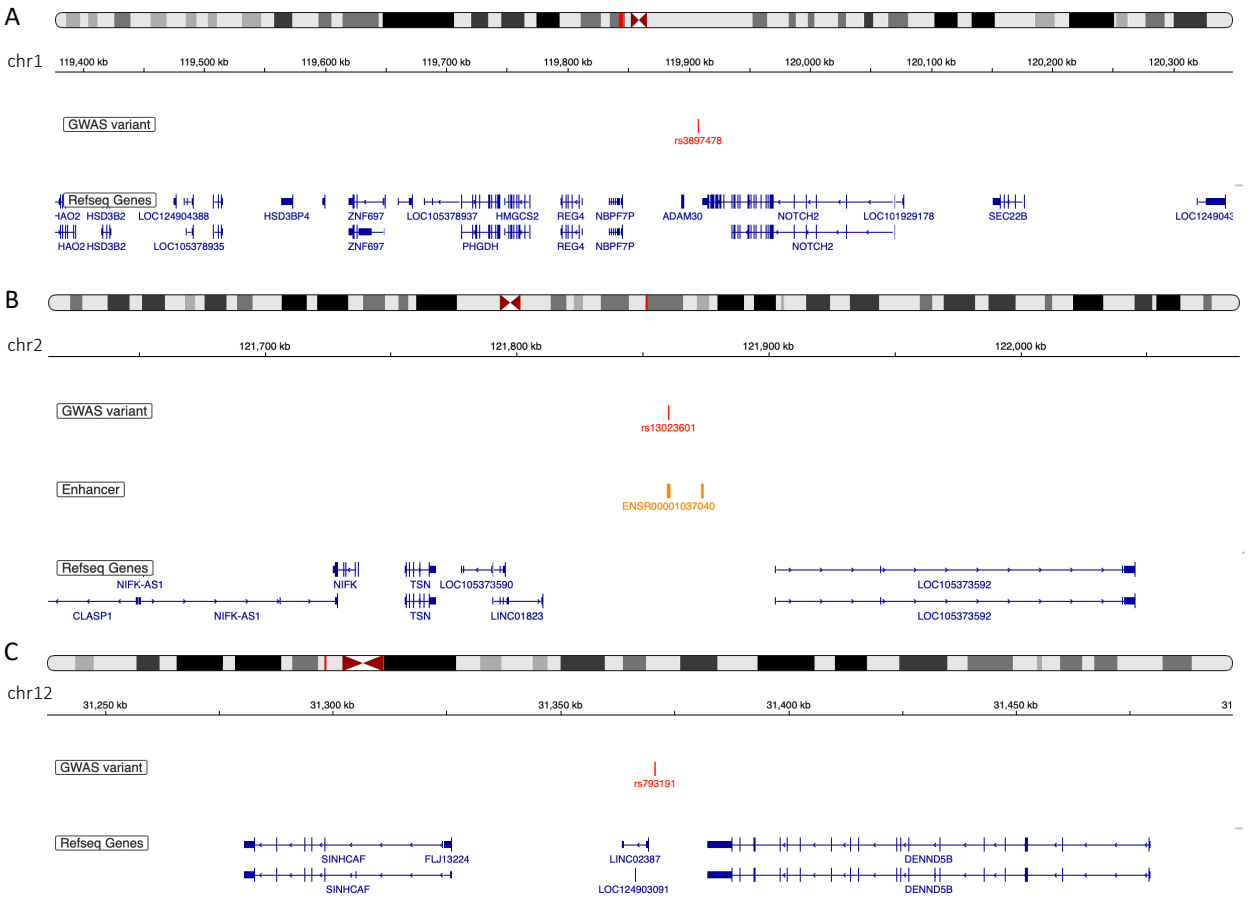



**Figure 4.13 Examples of GWAS variants predicted as with regulatory effects.** (A) Variant rs3897478 (chr1_119908567_T_C) and the corresponding gene track. The prediction probability of this variant being a regulatory variant is 0.90. (B) Variant rs13023601 (chr2_121860599_C_T) along with the corresponding enhancer and gene tracks. The prediction probability of the variant is 0.85. (C) Variant rs793191 (chr12_31370922_A_G) and the corresponding gene track. The prediction probability of the variant is 0.90.

### 4.3.3 Cattle regulatory variant prediction

The variation in performance among the human models trained using variants subjected to different filters highlights the significance of data quality in constructing reliable machine learning models. However, the fact that adequate accuracies were obtained even when using only features available for both species was promising for training cattle models. Considering that human genomic data is more dependable than cattle genomic data, we did explore the possibility of utilizing human data to enhance cattle regulatory

variant prediction. In this section, I initiated cattle regulatory variant prediction using human data and subsequently employed cattle annotations directly for prediction.

#### 4.3.3.1 Prediction results based on human annotations

Three strategies were investigated for employing human annotations to predict cattle regulatory variants: utilizing pre-trained human models for cattle regulatory variant prediction, employing incremental learning based on the human model for cattle prediction, and lifting cattle variants to the human genome to train the model using human annotations. The XGBoost, LightGBM, and CatBoost models were re-tuned using the new datasets and subsequently aggregated through a voting classifier in each strategy. The lifted-over cattle data (79215 foreground variants and 79215 background variants) were split into training and test sets at a ratio of 70% and 30%, respectively. Figure 4.14 shows the performance of these three strategies.

The model trained on the lifted-over data exhibited the best performance among these three strategies, achieving an AUROC of 0.708. The human model incremented with cattle lift-over data surpassed the performance of the model relying solely on human data, underscoring distinct regulatory variant patterns inherent to different species. The decrease in model performance when compared to human regulatory variant predictions could be attributed to several factors. Apart from the differences in data quality between regulatory variants defined in cattle GTEx and human GTEx, the models for cattle prediction encompass only a restricted set of annotations that are also available in cattle annotations. As for predicting regulatory variants in cattle using models trained on humans, matching features are required for the cattle variants.

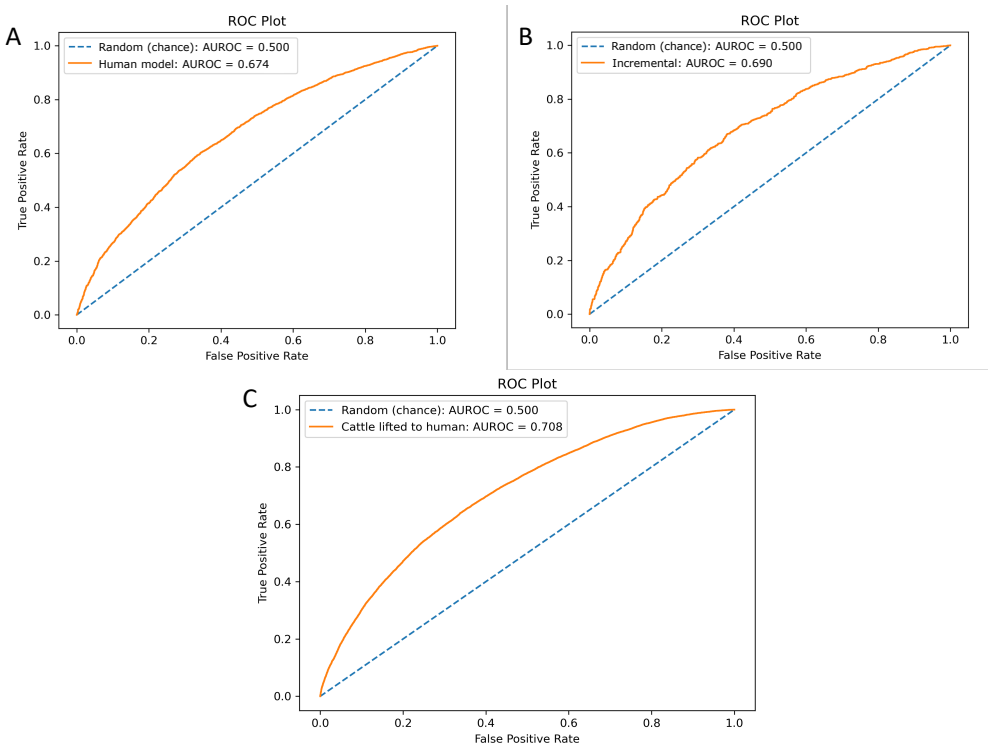



**Figure 4.14 ROC plots for different strategies of predicting cattle regulatory variants based on human annotations.** ROC plots and AUROC scores for (A) model trained using human data and tested on cattle variants. (B) model trained on human data and incremented on lift-over cattle data. (C) model trained and tested on lift-over cattle data.

### 4.3.3.2 Prediction results based on cattle annotations

Subsequently, I employed cattle annotations, rather than relying on human annotations, as features for predicting cattle regulatory variants. In addition to the model based on the entire dataset (158,430 variants), which encompasses all regulatory variants across different tissues, five tissue-specific models were trained and tested. The blood dataset included 16,328 variants, the muscle dataset included 5,378 variants, the mammary dataset included 10,376 variants, the liver dataset included 13,034 variants, and the adipose dataset included 6,308 variants. Each of these datasets were again divided into training and test sets using a ratio of 70% for training and 30% for testing.

**Table 4.6 Different metrics for cattle models**

| Tissue | Recall | Precision | Accuracy | AUROC | Average cv score [a] | Standard deviation of cv scores [b] |
|---|---|---|---|---|---|---|
| Blood | 0.663 | 0.665 | 0.660 | 0.725 | 0.664 | 0.002 |
| Liver | 0.670 | 0.656 | 0.655 | 0.719 | 0.662 | 0.002 |
| Adipose | 0.688 | 0.678 | 0.675 | 0.730 | 0.683 | 0.002 |
| Muscle | 0.658 | 0.673 | 0.656 | 0.716 | 0.665 | 0.004 |
| Mammary | 0.638 | 0.646 | 0.638 | 0.689 | 0.642 | 0.003 |
| Whole | 0.655 | 0.651 | 0.653 | 0.715 | 0.653 | 0.003 |

[a] *The average F1 score of the models obtained through 5-fold cross-validation.*
[b] *The standard deviation of the scores obtained from 5-fold cross-validation for the models*

In comparison to the model retrained on human annotations using cattle variants lifted to the human genome (AUROC=0.708), the model trained directly on cattle annotations attained a slightly elevated AUROC score of 0.715. According to a further check of the feature importance in these two models, a similar feature ranking was observed, with gene density and distance to CpG island consistently ranking as top features in both models (Figure S 4.1 and Figure S 4.2 ). Additionally, as indicated in Table 4.6, the tissue-specific models demonstrated superior performance compared to the model based on the entire dataset, except for the mammary-specific model. Among these tissue-specific models, the adipose-specific model exhibited the best performance across all metrics, despite comprising only 6,308 variants in this set. Next, I investigated the relationship between the nominal p-values and the prediction probabilities of the predicted regulatory variants in adipose and blood models (Figure 4.15). The nominal p-values were transformed to $-log_{10}^{nominal\ pval}$ for better clarity. There appears to be a weak but positive correlation between the prediction probabilities of the variants and the $-log_{10}^{nominal\ pval}$ from cattle GTEx in both models, indicating the potential utility of prediction probabilities from machine learning models as supplementary information for prioritizing regulatory variants in cattle tissues.

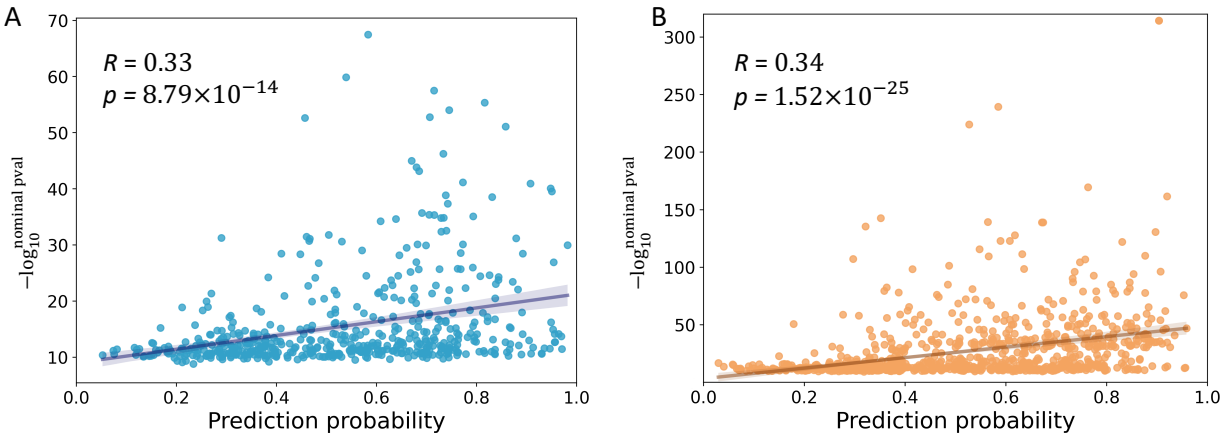



**Figure 4.15 The association between the variant prediction probabilities from the models and the variant nominal p-values from cattle GTEx.** (A) Prediction probabilities from the adipose-specific model. The line depicted on the plot represents the linear regression line fitted to the data points. (B) Prediction probabilities from the blood-specific model. The correlation coefficients and their corresponding p-values are specified in the plots.

To further assess the efficacy of the cattle models, I split the mammary, liver, blood, and the whole datasets into five distance-to-TSS bins: 0-5,000bp, 5,000bp-12,000bp, 12,000bp-30,000bp, 30,000bp-70,000bp, >70,000bp. Each bin was designed to contain a roughly equal number of variants within each dataset. The adipose and muscle datasets were excluded due to their relatively low variant counts. As depicted in Figure 4.16, the mammary model trained using the variants within 5,000bp distance to TSS outperformed both the blood and liver models trained on variants within the same range. It achieved the highest average cross validation score (F1 score) of 0.802, compared to 0.786 for blood and 0.790 for liver model. In general, these models performed better when restricted to variants near the TSS, indicating that predicting promoter regulatory variants is generally easier than predicting distal regulatory element variants.

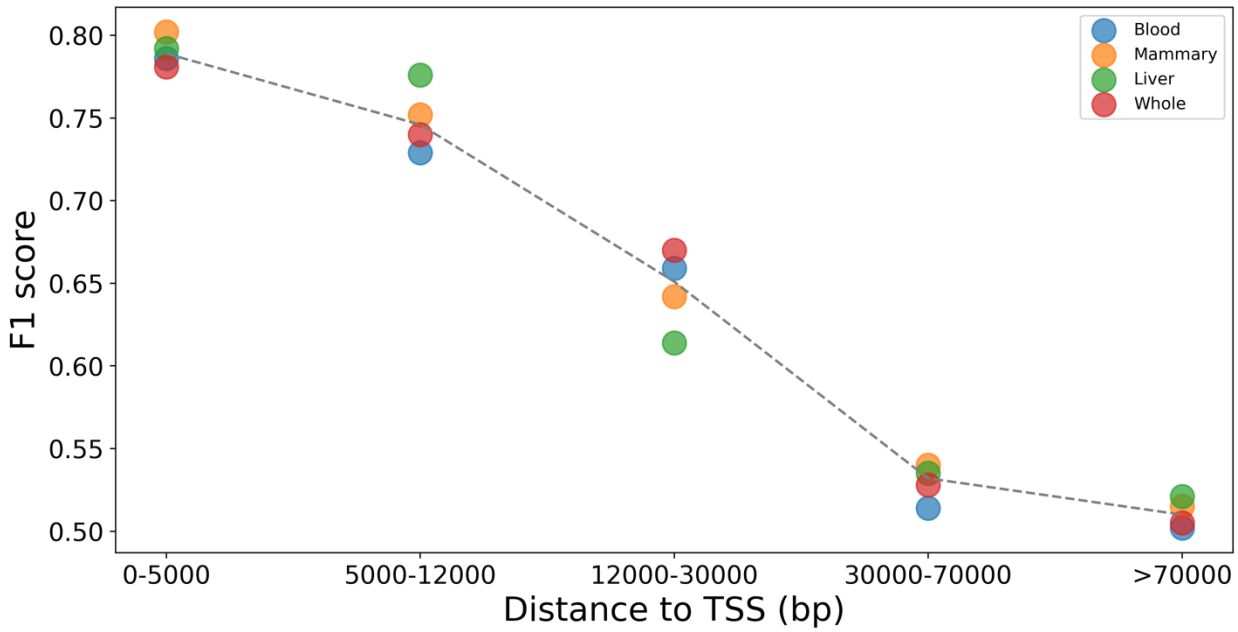



**Figure 4.16 Performance of cattle models when training and testing using variants within different distance ranges to the TSS.** The grey dashed line is plotted based on the median F1 score among the four datasets in each range.

Furthermore, the Enformer annotation workflow was employed to predict 5,313 chromatin and gene expression features for these five cattle tissues. Subsequently, this set of features underwent feature selection and was integrated with other cattle annotations to train the model for each tissue. However, the cattle tissue-specific models did not exhibit improvement with these additional Enformer features, achieving the AUROC scores of 0.720, 0.719, 0.728, 0.716, and 0.685 in blood, liver, adipose, muscle and mammary models, respectively. This suggests that the inclusion of features derived from the Enformer models trained on human chromatin and expression data do not improve the prediction of cattle regulatory variants.

### 4.3.4 Cattle SuRE regulatory variant prediction

Given the lower performance at predicting cattle regulatory variants using datasets from cattle GTEx, I proceeded to delve deeper into the model's performance by predicting cattle regulatory variants derived from the potentially more accurate MPRA approach. Whereas the cattle GTEx dataset is affected by issues such as incomplete genotyping,

due to relying on variants identified from RNA-seq data, collapsing of data across disparate studies, issues of LD, and variable sample sizes, these factors are largely mitigated with the SuRE data generated within my lab. However, an important caveat to note of the SuRE data is that each variant is tested without its native chromatin context, therefore the assay provides an indication of the regulatory potential of a variant that may not match its effect when in its endogenous location. The distribution of the initial SuRE regulatory variants across the chromosomes was investigated, revealing some enrichment at the telomeres of the chromosomes. Following the genotype quality (GQ) and sequencing depth (DP) filtering processes as described in section 4.2.2.2, the notable enrichment of variants at the telomeres was eliminated. After filtering, there were 18,782 variants retained as the foreground regulatory variants and the same number of background variants were randomly selected from the remaining data. The whole set of variants were annotated with the same features used in the cattle GTEx prediction. Then the voting classifier was trained on the training set and tested on a separate test set. Impressively, the raQTL model attained an AUROC score of 0.803 (Figure 4.17), surpassing the performance of any of the cattle eQTL models. Once again, this outcome underscores the substantial influence of selecting high-quality data on the performance of the machine learning model.

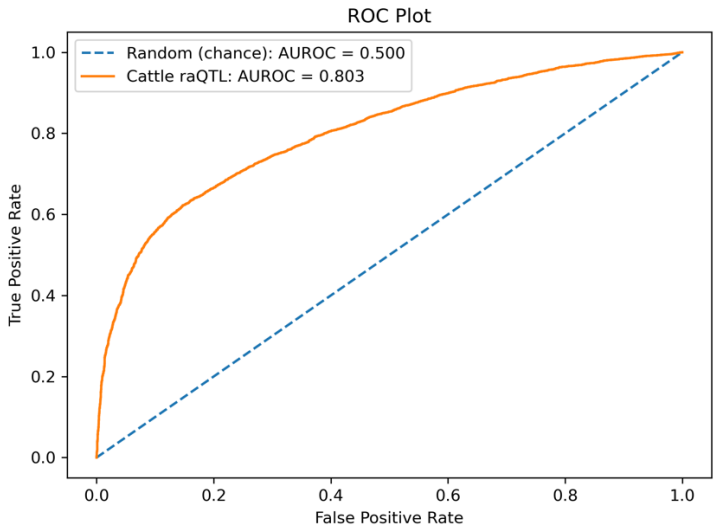
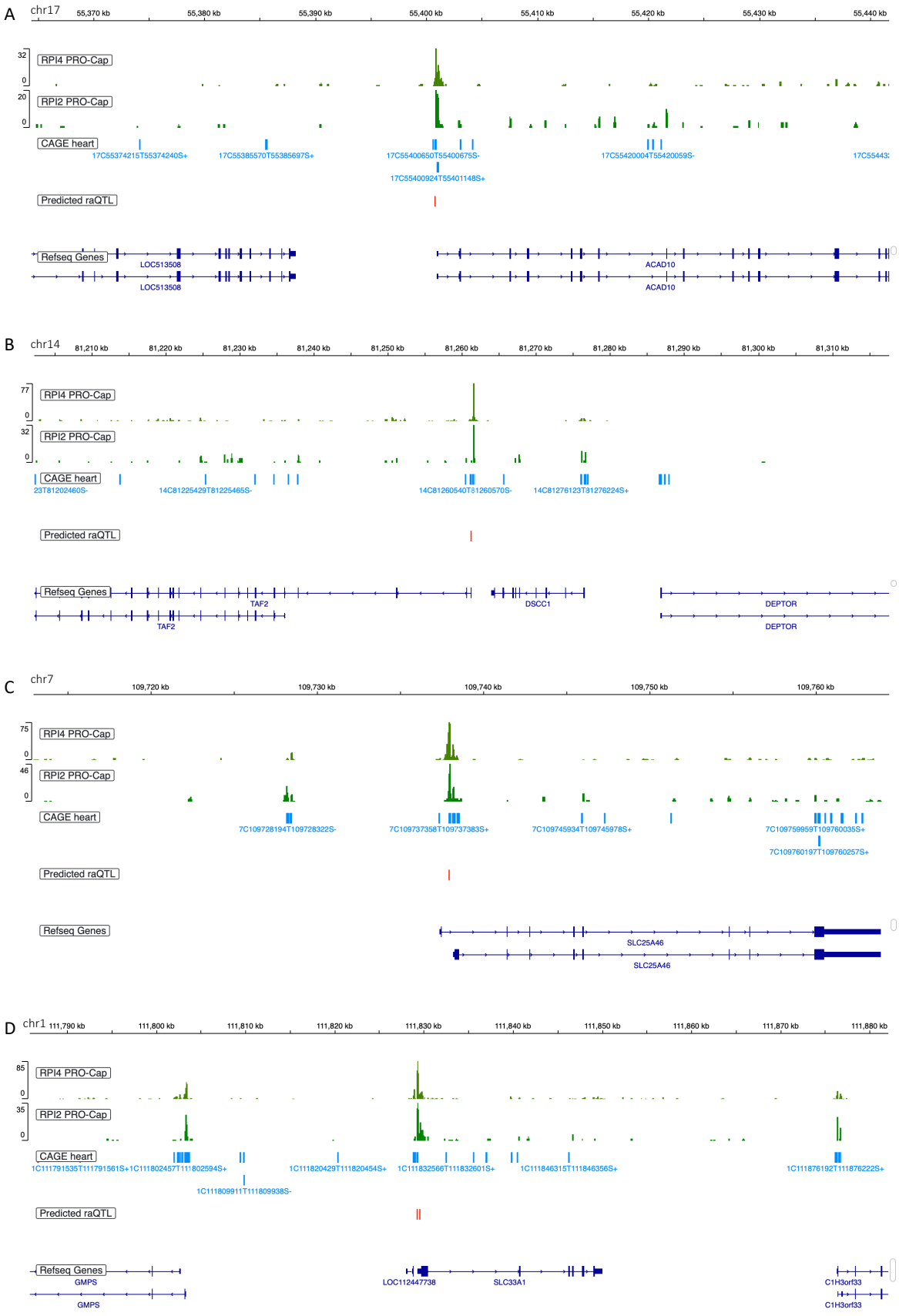



**Figure 4.17 ROC plot and AUROC score for the cattle SuRE regulatory variant model**

### 4.3.4.1 Examples of the predicted SuRE regulatory variants

To assess the potential underlying molecular mechanisms of the predicted SuRE regulatory variants, I conducted an analysis using the functional genomic data, focusing on variants falling within PRO-Cap peaks and TSS regions as measured by CAGE. Figure 4.18 shows some representative tracks from the Integrative Genomics Viewer (IGV), showcasing selected example regulatory variants with prediction probabilities greater than 0.9 from the machine learning model.

For instance, variant chr17:55400838 was identified within the TSS region of the *ACAD10* gene (Figure 4.18 A). Previous research (Bloom et al., 2020; Ghaffari et al., 2021) has highlighted the relevance of the *ACAD10* gene to fatty acid metabolism in various species, including cattle and mice. Variant chr14:81261310 was found in close proximity to the TSS region of an isoform of the *TAF2* gene, known for its significance in oocyte developmental potential in cattle (Walker & Biase, 2020). Another notable predicted regulatory variant chr7:109737970 was found located within the TSS region of one isoform of the *SLC25A46* gene, while situated within an intron region in another isoform. This gene has been found associated with mitochondrial dynamics and metabolism in both human and cattle (Duchesne et al., 2017). Furthermore, variant chr1:111829620 was positioned within the TSS region of the *SLC33A1 gene*, another member of the solute carrier family. These predicted regulatory variants, which are situated within or proximal to TSS regions of genes, may exert different effects on gene expression. Some could result in alternative start site usage or isoform-specific expression due to their appearance in certain isoforms of a gene, as seen with variants chr14:81261310 and chr7:109737970. Another predicted variant chr10:8012211 was found in the intergenic region as shown in Figure 4.18 E. This variant could potentially influence the distal regulatory elements of the *F2RL1* gene, which was found associated with heat-tolerance trait in cattle (Cheruiyot et al., 2021).


A
chr17
55,370 kb
55,380 kb
55,390 kb
55,400 kb
55,410 kb
55,420 kb
55,430 kb
55,440 kb
RPI4 PRO-Cap
RPI2 PRO-Cap
CAGE heart
17C55374215T55374240S+
17C55385570T55385697S+
17C55400650T55400675S-
17C55420004T55420059S-
17C55400924T55401148S+
Predicted raQTL
Refseq Genes
LOC513508
ACAD10
B
chr14
81,210 kb
81,220 kb
81,230 kb
81,240 kb
81,250 kb
81,260 kb
81,270 kb
81,280 kb
81,290 kb
81,300 kb
81,310 kb
RPI4 PRO-Cap
RPI2 PRO-Cap
CAGE heart
23T81202460S-
14C81225429T81225465S-
14C81260540T81260570S-
14C81276123T81276224S+
Predicted raQTL
Refseq Genes
TAF2
DSCC1
DEPTOR
C
chr7
109,720 kb
109,730 kb
109,740 kb
109,750 kb
109,760 kb
RPI4 PRO-Cap
RPI2 PRO-Cap
CAGE heart
7C109728194T109728322S-
7C109737358T109737383S+
7C109745934T109745978S+
7C109759959T109760035S+
7C109760197T109760257S+
Predicted raQTL
Refseq Genes
SLC25A46
D
chr1
111,790 kb
111,800 kb
111,810 kb
111,820 kb
111,830 kb
111,840 kb
111,850 kb
111,860 kb
111,870 kb
111,880 kb
RPI4 PRO-Cap
RPI2 PRO-Cap
CAGE heart
1C111791535T111791561S+
1C111802457T111802594S+
1C111809911T111809938S-
1C111820429T111820454S+
1C111832566T111832601S+
1C111846315T111846356S+
1C111876192T111876222S+
Predicted raQTL
Refseq Genes
GMPS
LOC112447738
SLC33A1
C1H3orf33

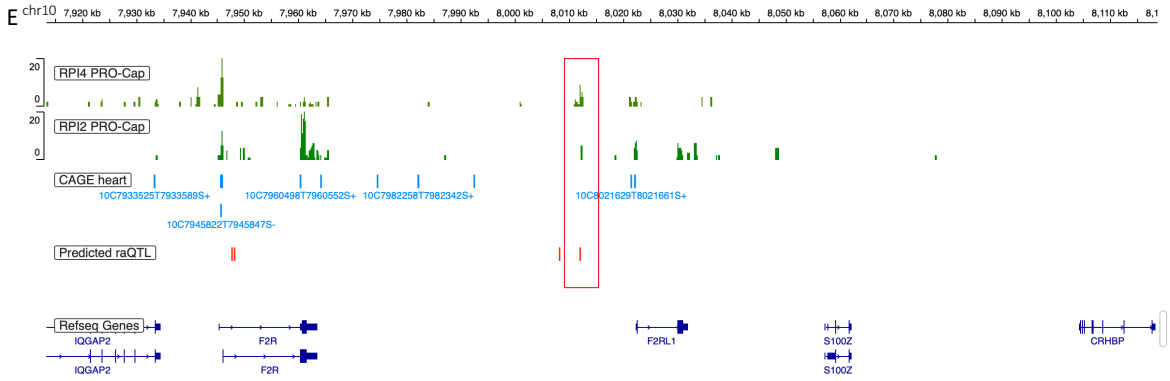



**Figure 4.18 Representative IGV tracks showing the locations of PRO-Cap and CAGE peaks, and example variants falling within these regions.** (A) Tracks for predicted regulatory variant chr17:55400838 with a prediction probability of 0.91. The green tracks are the PRO-Cap tracks in two samples. The blue track shows the CAGE-measured TSS regions. The target variant is represented in red. The dark blue track is the gene track in the region. (B) Tracks for the predicted regulatory variant chr14:81261310 with a prediction probability of 0.95. (C) Tracks for the predicted regulatory variant chr7:109737970 with a prediction probability of 0.94. (D) Tracks for the predicted regulatory variant chr1:111829620 with a prediction probability of 0.98. (E) Tracks for the predicted regulatory variant chr10:8012211 with a prediction probability of 0.98.

## 4.4 Discussion

The identification of regulatory variants remains a challenge primarily due to their enrichment in non-coding regions and the presence of LD patterns among adjacent variants. This challenge is especially pronounced in livestock species, where the availability of high-quality functional genomic data is comparatively limited. To tackle this issue, this chapter delved into the potential of utilizing machine learning approaches for predicting regulatory variants in both humans and cattle.

Before model training, the distribution disparities in annotations between regulatory variants and other variants were investigated in both human and cattle. Certain characteristics associated with regulatory variants were consistently identified across both species, including enrichment in gene-dense regions and specific genomic elements, as well as an increased presence of C:G pairs within their flanking sequences. However,

a notable difference was observed in the distribution of conservation scores between human and cattle. In humans, eQTLs displayed a slight enrichment near a phastCons30way score of 0, signifying their preference for less constrained genomic regions, which was not as expected. However, a recent study by Mostafavi et al. revealed that genes situated near eQTLs exhibit a scarcity of functional annotations, display a reduced level of constraint, and possess less complex regulatory landscapes compared to the GWAS hits in humans (Mostafavi et al., 2022). This finding could potentially elucidate the observed slight enrichment of human eQTLs in regions characterized by low phastCons scores. On the other hand, cattle eQTLs showed a reduced enrichment near a phastCons241way score of 0, but demonstrated an enrichment in regions with higher phastCons241way score. It should be noted that human and cattle conservation scores were calculated based on different multiple alignments. The human phastCons30way conservation score was based on a 30-species multiple alignment with a focus on primates (26 out of 30), while the cattle phastCons241way conservation score was based on a 241-species multiple alignment, including a much more diverse range of species. The varying number of species and their distinct evolutionary time scales could potentially account for the differences in conservation scores. Therefore, it may be worthwhile to generate human conservation scores using the 241-way multiple alignment and further compare the conservation differences across species.

The prediction results in both human and cattle underscore the significance of data quality in influencing model performance. In human, liver-specific model achieved the highest AUROC score of 0.919, showcasing better overall performance across various distances from the TSS among five different tissues. The human liver is known to have a relatively high number of tissue-specific eQTLs, attributed to a number of genes being associated with liver-specific expression patterns (Glubb et al., 2012). Importantly liver is a relatively homogenous tissue likely making the accurate definition of regulatory variants easier, compared to other tissues that are more confounded by cell type heterogeneity. The distinct characteristics of these liver-specific eQTLs contribute to the model's ability to capture informative features, thereby leading to a commendable proficiency in classifying previously unseen variants. In cattle GTEx regulatory variant prediction, the adipose

model exhibited the highest AUROC of 0.730 among the five different tissues, indicating that eQTLs in cattle adipose may possess greater specificity compared to the other four tissues. Furthermore, the model trained using regulatory variants in aortic endothelial cells based on MPRA approach (AUROC=0.803) outperformed all models trained on cattle GTEx data, suggesting potentially higher reliability of regulatory variants based on MPRA approaches. It is important to note that the MPRA and cattle GTEx approaches utilize different filters, leading to distinct sets of tested variants. The MPRA model is trained and tested on variants known to be expressed, while the cattle GTEx models encompass a wider range, including both expressed and unexpressed ones. This difference in variant sets is likely a factor contributing to the performance difference between the models. Currently, the MPRA model is trained and tested using data from a single cell line. To provide a more compelling case for the superiority of MPRA models compared to cattle GTEx models, future studies could involve training and testing MPRA models on data from a broader range of tissues/cell lines. Demonstrating consistently stronger performance across diverse cell lines would strengthen the argument that MPRA models offer a generally better approach for regulatory variant prediction in cattle. Furthermore, the cattle model based on raQTLs still exhibited lower performance compared to the human models trained using features available across both species (Table 4.5). This difference could be attributed to multiple factors, including variations in genomic data quality and differences in the complexity of regulatory landscapes between the two species.

In cattle GTEx regulatory variant prediction, two approaches were investigated, including the one based on human annotations and the other one based on cattle annotations. Figure 4.19 shows the performance ranking of the different cattle models based on their AUROC scores. The model trained with cattle annotations (AUROC=0.715) exhibited slightly superior performance compared to the human-based cattle model (AUROC=0.708). This outcome likely reflects to some extent divergence between humans and cattle, resulting in variations within the regulatory landscape, including disparities in gene expression patterns and regulatory mechanisms. Although the human-based model captured some general regulatory principles shared across both species, it

might overlook species-specific intricacies necessary for enhanced prediction accuracy. This was supported by the improved performance of the incremental model, which utilized cattle lift-over data to retrain the human-based model, when compared to the model solely based on human annotations.

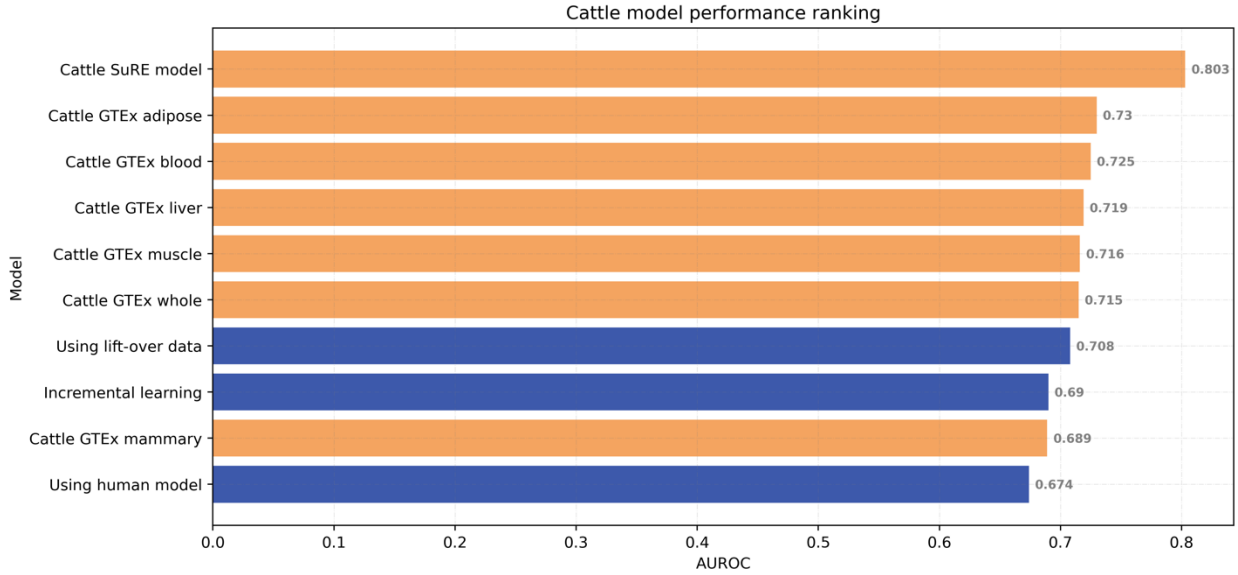



**Figure 4.19 Cattle models performance ranking.** The models were ranked in descending order according to their AUROC scores. The orange bar refers to the model using cattle annotations, while the blue bar indicates the models trained based on human annotations.

Another notable observation from the study pertains to the limitations of the deep learning-based architecture for gene expression level prediction in predicting regulatory variants. Specifically, the human model trained on the general features (AUROC=0.885) outperformed the model trained with Enformer features (AUROC=0.843) and exhibited similar performance to the model trained on the combined feature set encompassing both general and Enformer features (AUROC=0.884). Moreover, the inclusion of Enformer features did not lead to improved performance in the cattle model either. This observation aligns with a recent study that underscored the deficiencies of various genomic deep learning models in elucidating gene expression variation (Huang et al., 2023).

In conclusion, this chapter has demonstrated the efficacy of employing machine learning techniques to predict regulatory variants in both human and cattle. The general annotations from the variant annotation pipeline proposed in Chapter 2 were proven to be effective as features in machine learning models for regulatory variants prediction, resulting in the highest AUROC score of 0.919 in the human liver-specific model. For cattle regulatory variants, the model based on cattle annotations exhibited a slightly superior performance in comparison to the one utilizing human annotations, achieving an AUROC score of 0.803 in the model trained using regulatory variants based on MPRA approaches.

These models can be further utilized for prioritizing variants associated with cattle traits of interest for subsequent editing. Moreover, they have the potential to aid in the prioritization of variants for calculating Genomic Estimated Breeding Values (gEBVs). The generation of gEBVs doesn't necessarily require the identification of causal variants; instead, it can be achieved by considering variants in LD with the causal variants (Xiang et al., 2021). Therefore, compared to the prioritization of variants for genome editing, the application of the models in predicting gEBVs may not require exceptionally high prediction accuracy. Enriching the dataset with functional variants is likely to improve the predictions.

# 4.5 Supplementary material

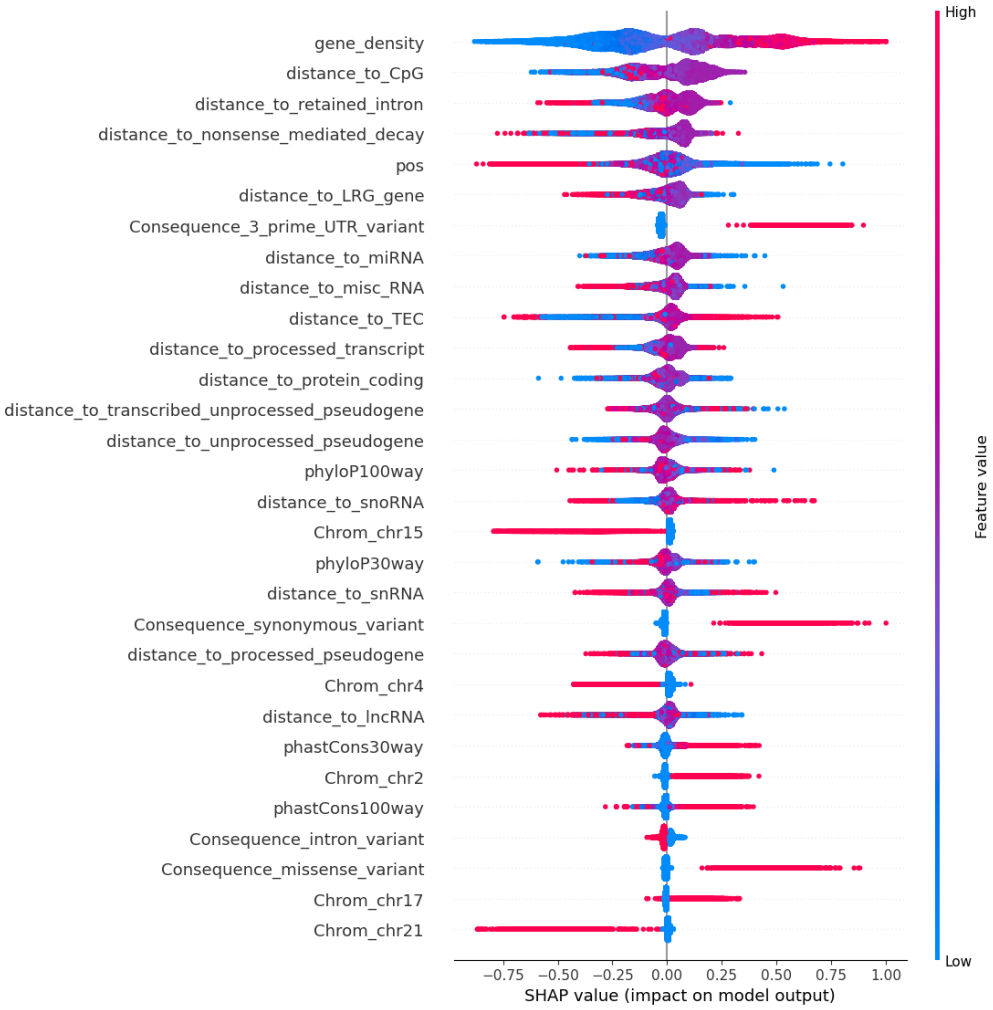



**Figure S4.1 SHAP plot for top 30 features in model based on lift-over cattle data.**

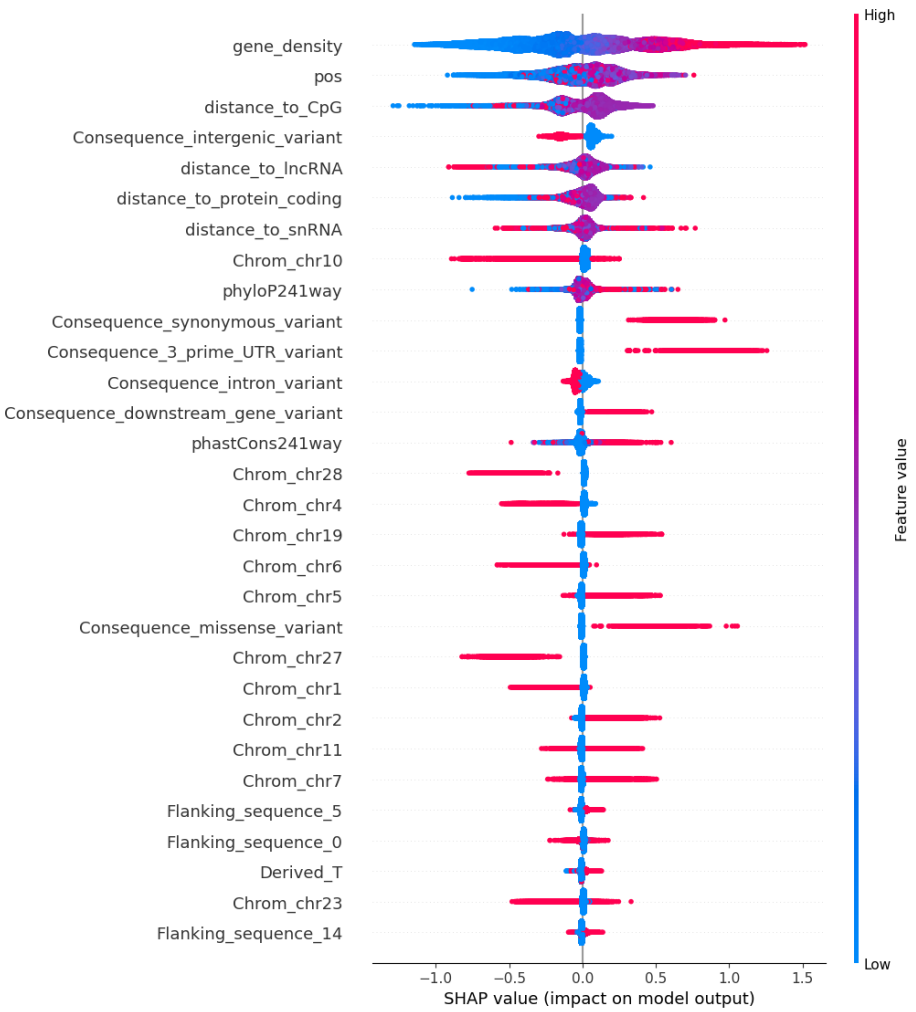



**Figure S4.2 SHAP plot for top 30 features in model based on cattle annotations.**

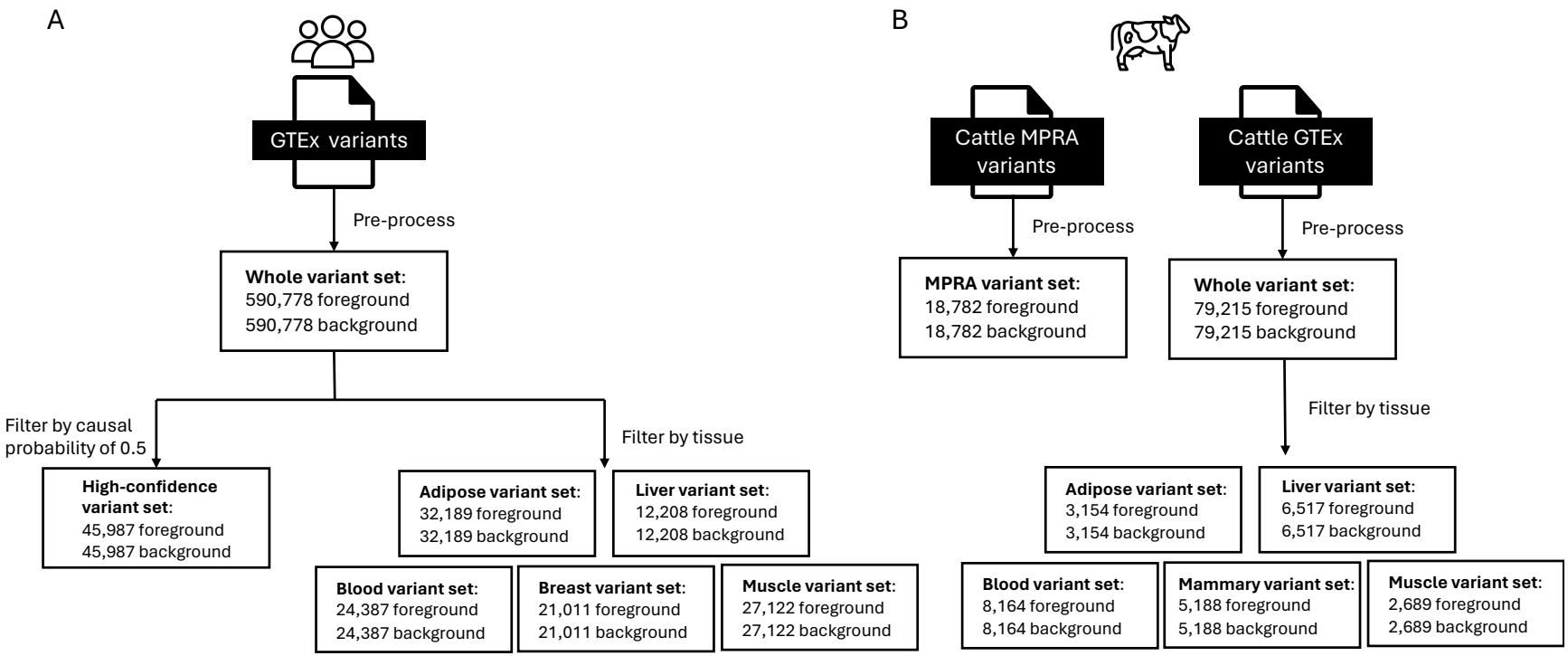



**Figure S4.3 Summary of different variant sets used in human** (A) **and cattle** (B) **modelling work.**

**Table S4.1 Summary of human regulatory variants prediction work**

| Species | Chapter | experiment | Machine learning algorithm | Variant source | Feature set |
|---|---|---|---|---|---|
| Human | 4.3.2.1 Prediction results for regulatory variants across different tissues | Baseline model test | Random Forest | Human GTEx | General features as specified in Table 4.1 |
| | | Machine learning algorithms comparison | Random Forest | | |
| | | | CatBoost | | |
| | | | XGBoost | | |
| | | | LightGBM | | |
| | | | GBDT | | |
| | | | SVM | | |
| | | | Voting (hard) | | |
| | | | Voting (soft) | | |
| | 4.3.2.2 Comparison of model performance based on different feature sets | Enformer | CatBoost | | Enformer features |
| | | Enformer + general features | CatBoost | | General features |
| | | Selected Enformer features + general features | Voting (soft) | | Selected Enformer features + general features |
| | | EpiMap + general features | Voting (soft) | | EpiMap + general features |
| | 4.3.2.3 Prediction results for tissue-specific data | Regulatory variants prediction in five tissues (blood, liver, adipose, muscle and breast) | Voting (soft) | | General features |
| | | | | | Features available in both cattle and humans |

**Table S4.2 Summary of cattle regulatory variants prediction work**

| Species | Chapter | experiment | Machine learning algorithm | Variant source | Feature set |
|---|---|---|---|---|---|
| Cattle | 4.3.2.1 Prediction results based on human annotations | Using human model | Voting (soft) | Cattle GTEx | Human features available in both cattle and humans |
| | | Incremental learning based on human model | | | |
| | | Lifting cattle variants to human | | | |
| | 4.3.2.2 Prediction results based on cattle annotations | Using entire dataset across all tissues | | | Cattle features |
| | | Tissue-specific (blood, liver, adipose, muscle, mammary) | | | Cattle features |
| | | | | | Cattle features + selected Enformer features |
| | 4.3.3 Cattle SuRE regulatory variants prediction | SuRE regulatory variants prediction | | SuRE | Cattle features |

# Chapter 5 Final discussion

Functional variants are important drivers of diverse diseases and phenotypic traits across both human and livestock species. Unlike humans, where there is a substantial body of research on functional variants, there is far less information available for livestock species. Studying functional variants in livestock species is essential for downstream research endeavors, including selective breeding, improving animal welfare, and increasing agricultural efficiency. However, the identification of functional variants remains a challenge primarily attributed to intricate genetic phenomena, most notably linkage disequilibrium (LD), confounding the precise identification of causal variants due to their non-random associations with neighbouring genetic markers. The application of machine learning approaches, with their inherent ability for pattern recognition and predictive modelling from large datasets, presents a promising avenue to unravel these complex relationships and improve the prediction of functional variants.

In this thesis, I have investigated the effectiveness of machine learning approaches in predicting functional variants in mammals. I first developed a reusable variant annotation pipeline, which provides a comprehensive set of annotations, ranging from variant position properties to sequence conservation, for both human and other mammalian species (Chapter 2). These informative annotations have subsequently served as valuable features for machine learning models employed in this study. Following the variant annotation pipeline, I initially applied it to predict whether human variants have livestock orthologues, with the aim of exploring the conservation of human functional variants in livestock species (Chapter 3). An extensive sharing of variants was found across human and livestock species, including hundreds of conserved functional variants. Subsequently, I employed the annotations to predict regulatory variants in both humans and cattle (Chapter 4). I conducted various experiments using different variant sets, diverse feature sets, and different machine learning algorithms. In human predictions, the liver-specific model outperformed all other human models and achieved an AUROC score of 0.919. Given the genomic conservation between human and cattle, as well as the higher reliability of human data, I initiated cattle regulatory variant prediction based on

human data and compared the results with models based on cattle annotations directly. The model trained on the MPRA regulatory variants using cattle annotations achieved the best cattle model performance with an AUROC of 0.803.

In this chapter, I will start by summarizing the work and results in this thesis, and then discuss the future directions for utilizing machine learning in predicting functional variants in mammals.

## 5.1 Summary of work and results

### 5.1.1 Variant annotations for machine learning applications

To date, numerous variant annotation tools have been developed based on various genomic datasets and databases. However, these tools tend to exhibit biases towards humans and are primarily designed for more generic applications. To address the issue of the lack of variant annotation tools for livestock species, especially tailored for downstream machine learning applications, I have developed a reusable variant annotation pipeline that can be applied across different species (Chapter 2). The pipeline offers five categories of annotations for human and other mammalian species: sequence conservation, variant position properties, VEP annotation, sequence context, and Enformer scores. These annotations provide abundant information about the variants, which are valuable for downstream machine learning models.

For the first time, I generated conservation scores for cattle and developed the workflow for calculating phastCons and phyloP conservation scores for other livestock species based on the Cactus multiple alignment, which includes 241 different species. The distribution of these two conservation scores in coding and non-coding regions across the cattle genome was evaluated in Chapter 4, with the coding regions showing higher conservation compared to the non-coding regions. This validates the dependability of my approach for generating novel conservation scores for livestock species. Evolutionary constraint is an important resource for exploring the functional importance of the genomic positions, thereby aiding in the understanding of genetic variations and their downstream

impact. Previous research has predominantly focused on human, primates, and rodents (Nassar et al., 2023), resulting in a lack of conservation resources in livestock species. The workflow proposed in the thesis offers an opportunity to obtain conservation scores for different livestock species included in the 241-way multiple alignment. These scores can be utilized not only in machine learning approaches but also in various applications, such as improving functionally informed fine-mapping analyses in livestock species. The recent Zoonomia project has conducted comprehensive research on utilizing the human conservation scores derived from the same 241-way Cactus alignment to evaluate human variants, genes and diseases (Sullivan et al., 2023). These approaches can be extended to livestock species by utilizing the conservation scores generated from the pipeline, thus improving our understanding of livestock species diseases and phenotypes.

### 5.1.2 The conservation of human functional variants across livestock species

Before predicting functional variants in specific species, I first employed the annotations as features to investigate the potential of predicting human variants that are conserved across livestock species (Chapter 3). This study aims to explore the extent to which human functional variants are conserved across livestock species. It sought to provide insights into the feasibility of leveraging the extensive research on human functional variants for less well-studied livestock species by considering naturally occurring functional variants shared across species. Additionally, it aims to explore the potential of utilizing livestock species with orthologous functional variants as naturally occurring animal models for studying human diseases.

Over 1.1 million human SNPs were found to have a directly orthologous variant in either cattle or pig cohorts. After analysing the distribution difference of the annotations of human variants with or without livestock orthologues, several features were found associated with those with livestock orthologues, including sequence context, distance from the variants to TSS and other genomic elements. Subsequently, machine learning models were trained using these annotations as features to predict whether human variants have livestock orthologues. Sequence conservation emerges as the most

important feature guiding the prediction process, with human variants situated in less-conserved regions exhibiting a higher likelihood of having an orthologue in another species. In these less-conserved regions, the selection pressure is often weaker or even neutral compared to conserved regions. This lower selection pressure allows variations in less-conserved regions to persist, thereby increasing the probability of their appearance in other species.

Following the investigation of the conservation of general human variants across species, we investigated the conservation of human functional variants across livestock species. Several hundred human pathogenic variants were discovered to be conserved across livestock species, including some associated with important diseases such as cancers. This collection of shared pathogenic variants offers opportunities for subsequent research into these diseases using naturally occurring animal models. Furthermore, functional variants associated with specific traits, such as biotinidase deficiency and height, were found more likely to have orthologues in livestock species such as cattle. These variants may have been conserved across species because they were selected for by humans. These observations not only yield insights into the evolutionary trajectories of these livestock species, but also identify prospective candidate variants that could be considered for integration into livestock breeding programs to improve livestock traits.

By comparing the fine-mapped regulatory variants from the human and cattle GTEx projects, 221 human regulatory variants were found have corresponding cattle regulatory variants tested on the orthologous genes in at least one tissue. Notably, these orthologous regulatory variants commonly exhibit conserved directions of effect across the two species, implying that the investigation of their downstream impacts can be effectively conducted in these respective species. Moreover, the examination of conserved functional variants across different species serves to improve the study of their prospective roles by increasing the sample sizes, a particular advantage when analyzing rare functional variants.

Subsequent to our findings and their associated implications, recent studies have undertaken investigations in specific fields. Bertram et al. conducted an investigation of coincident SNPs in chicken and duck that lead to their different immune-related responses to avian influenza virus (AIV) (Bertram et al., 2023). Li et al. built a deep learning-based architecture DeepGCF to learn the functional conservation between human and pig, and further validated the conservation of regulatory variants between these two species by comparing the DeepGCF scores between the orthologous eQTLs and the background genome (J. Li et al., 2023).

While there exist hundreds of conserved functional variants across species that hold promise for transferring research insights from human functional variants to livestock species, particularly those related to breeding values, they constitute only a small fraction of all functional variants in livestock species. Therefore, in Chapter 4, I delved into the utilization of machine learning approaches to predict regulatory variants within specific species, aiming to enhance the initial detection of novel functional variants in livestock species.

### 5.1.3 The utility of machine learning approaches in regulatory variant prediction across mammalian species

The variant annotations and machine learning approaches were initially applied to predict human regulatory variants, as humans possess the most reliable and abundant data and provides an upper-bound of the likely performance of any modelling approach (Chapter 4). Consequently the prediction results in humans can serve as a reference for assessing the optimal performance achieved by the models across different species. Subsequently, I explored various approaches for predicting regulatory variants in cattle, including those based on human annotations and those relying solely on cattle annotations. This section summarizes the results of Chapter 4.

### 5.1.3.1 Tissue-specific models achieve a generally better performance

Considering that regulatory variants are tissue-specific, the machine learning models were trained using both entire datasets of fine-mapped variants as well as tissue-specific datasets. In human predictions, models based on tissue-specific variants without filtering by CaVEMaN prediction probability achieved comparable or better performance compared to the model trained on the entire high-confidence data (CaVEMaN prediction probability > 0.5). Notably, the liver-specific model achieved the best performance with an AUROC score of 0.919. The excellent performance of the liver-specific model may be attributed to the liver's relative cellular homogeneity compared to other tissues, enabling it to capture distinct characteristics useful for distinguishing regulatory variants. In cattle GTEx regulatory variant prediction, the tissue-specific models also achieved generally better performance compared to the entire dataset. These observations highlight the importance of considering tissue context when predicting regulatory variants, as well as the importance of data quality in machine learning approaches.

### 5.1.3.2 Models perform better in predicting regulatory variants near TSS

Both the human and cattle models, whether tissue-specific models or models based on the entire dataset, were found to perform better in predicting regulatory variants near the TSS than predicting those located farther from the TSS. This can be attributed to the complexity of long-range regulatory interactions, which usually involves the three-dimensional folding of chromatin, wherein regulatory elements establish physical contacts with their target genes (Dekker & Mirny, 2016). More importantly, the sequence rules in the regions near TSSs are more pronounced, whereas the impact of distal elements is weaker and more variable. Machine learning models may encounter difficulties in capturing these intricate interactions and complex rules to make predictions. Notably, the human liver-specific model consistently displayed good performance across different distance ranges to TSS, even when training on variants located farther away from TSS (> 100kb). This suggests the liver-specific model's proficiency in predicting regulatory variants not only within promoter regions but also those associated with distal regulatory

elements, indicating the potential existence of unique long-range regulatory mechanisms specific to liver that can be effectively captured by machine learning models.

#### 5.1.3.3 Cattle models based on cattle annotations slightly outperform human-based models

In cattle regulatory predictions, different approaches were explored, including both incorporating human data or based solely on cattle annotations. The cattle GTEx prediction results show that model using cattle annotations (AUROC=0.715) slightly outperformed the model trained based on human data (AUROC=0.708), indicating the limitation of leveraging human-based models for regulatory variant prediction in livestock species.

In the cattle annotations-based models, the one trained using regulatory variants in the aortic endothelial cells obtained from MPRA approach achieved the highest AUROC score of 0.803, in contrast to the best-performing cattle GTEx adipose model with an AUROC of 0.730. This indicates that the cattle regulatory variants generated through MPRA approaches are potentially more reliable than those from cattle GTEx. However, MPRA approaches do have some limitations, one of the major limitations is the isolated exogenous environment, which may not fully capture the complexity of the endogenous regulatory regions (Mulvey et al., 2021). Consequently, MPRA approaches has the potential to generate false-positive outcomes since not all observed impacts on gene expression hold functional relevance or biological significance. These limitations, as well as the size of training data, and genomic data quality, leading to the lower performance compared to human models trained using features available across both species.

Going forward, the cattle model can be further utilized to prioritize functional variants linked to traits, particularly useful for aiding in the prioritization of variants for calculating Genomic Estimated Breeding Values (gEBVs). Unlike the direct identification of functional variants, prioritizing variants for gEBVs calculation does not necessarily require the precise identification of the causal variants, as long as those variants in LD with the causal ones are detected (Xiang et al., 2021). Hence, the utilization of these models in predicting

gEBVs may not demand exceptionally high prediction accuracies. Enriching the dataset with functional variants is likely to improve the prediction of gEBVs (Xiang et al., 2021).

#### 5.1.3.4 Enformer features have limited contribution to model performance

To date, deep learning-based models, especially those popular models in natural language processing (NLP), have been applied to understand DNA sequences, thanks to their ability in capturing information from the sequence data. In this study, in addition to the general features from my main annotation workflow, I also explored the utility of the deep learning architecture Enformer in predicting regulatory variants in both human and cattle. Enformer was proved to be effective in predicting gene expression from the DNA sequence, achieving a Pearson correlation of 0.85 (Avsec et al., 2021). However, the prediction results in my study illustrate that Enformer features did not provide additional discriminative information beyond my general features for both human and cattle models. This observation aligns with a recent study that evaluated four popular deep learning architectures, including Enformer, Basenji2, Xpresso, and Expecto, in predicting gene expression variations (Huang et al., 2023). All these models showed limited performance in elucidating the impact of single base changes on gene expression.

Another recent study conducted *in silico* investigations to assess the performance of sequence-based models, primarily focusing on Enformer, in capturing gene expression in both promoter regions and distal enhancers in humans (Karollus et al., 2023). Through their experiments, they observed that Enformer is good at capturing causal factors of promoters but encounter difficulties when attempting to capture the causal impact of distal enhancers on gene expression. Despite Enformer's ability to accept sequences of up to 196kb as inputs, it does not fully utilize all the information within the long range. Removing two-thirds of the input has only a very small effect on the final predictions. Based on these findings, they also proved that it is rare to obtain meaningful outcomes when utilizing Enformer to predict the influence of distal variants on gene expression. These observations can further support and explain the limitations of utilizing Enformer features

in my study for predicting regulatory variants, despite the proposed promise of such models for this purpose.

## 5.2 Future directions

This work has explored the potential of utilizing machine learning approaches to predict functional variants, ranging from predicting functional variants shared across human and livestock species to predicting regulatory variants in specific species, especially in livestock species. The machine learning approaches, as well as the variant annotation pipeline, can be extended to other mammalian species. However, as discussed in the previous chapters, much is yet to be improved and explored. This section will summarize potential future directions in this field.

### 5.2.1 Extending the work to other types of variants

The annotation pipeline developed in this thesis has a focus on annotating single nucleotide variants. However, as introduced in Chapter 1, other types of sequence variants and structural variants (SVs) also play a crucial role in the development of genetic diseases and traits. Therefore, it is necessary to extend the variant annotation pipeline to other types of variants. These machine learning-targeted annotations can be further utilized in the machine learning approaches for understanding the impact of different types of variants. Some recent studies have annotated human structural variants with different features and try to predict their pathogenicity. For example, Sharo et al. developed StrVCTVRE, a Random Forest-based model to predict the pathogenicity of human SVs (Sharo et al., 2022). They annotated each SV using 17 features including conservation scores, gene importance, and exon expression. The model achieved an AUC score of 0.91 on large SVs (length greater than 500kb), an AUC score of 0.80 on mid-length SVs (30kb < SVs < 500kb), and an AUC of 0.89 on small SVs. Another application utilized three groups of features, including conservation scores, functional genomics data, and annotation metrics, to annotate SVs and built a machine learning-based architecture to calculate pathogenicity scores for these SVs (Kumar et al., 2020).

However, most current research has concentrated on humans, the impact of SVs in livestock species remains underexplored. Furthermore, their annotations have been generated from various sources, lacking a unified pipeline. Therefore, there is a potential to further extend the current pipeline to include structural variant annotation across livestock species. This expansion presents opportunities for employing machine learning approaches to better understand the potential impact of SVs in livestock species.

### 5.2.2 High quality data in livestock species

The difference in model performance between human and cattle, even when using shared features, highlights the unsatisfied data quality in cattle. Compared to humans, cattle lack abundant annotation resources and highly reliable variant data. Although I have utilized the regulatory variants based on MPRA approach as training and testing data, the reliability of these variants is still limited due to the inherent constraints of MPRA approaches. Furthermore, the MPRA data is derived from only one cell line, and the number of variants is not as abundant as in humans. Therefore, a comprehensive collection of high-quality cattle data, as well as data for other livestock species, encompassing functional genomic data and high-quality functional variants, is indispensable. The performance of livestock models in prioritizing novel functional variants can be significantly improved with the availability of more dependable training data.

### 5.2.3 Enhancing the prediction of functional variants associated with distant regulatory elements

The decrease in the models' performance when using variants farther away from the TSS in both human and cattle indicates the limitation of current machine learning models in predicting regulatory variants associated with distal regulatory elements. One of the major reasons for this is the complexity of long-range regulatory interactions. As discussed previously, it is challenging to capture the impact of distal regulatory elements on gene expression, even using the state-of-the-art deep learning architectures like Enformer.

Without accurate prediction of the impact of distal regulatory elements, it becomes difficult to predict functional variants associated with these elements.

There are some potential approaches to improve the performance of these deep learning-based architectures in capturing the impact of distal regulatory elements. The most straightforward way is to include more epigenetic and gene expression data related to distal regulatory element activities from more species and more cell lines (Karollus et al., 2023). An alternative strategy involves the integration of Enformer-like models with those specialized in predicting three-dimensional DNA contacts (de Almeida et al., 2022; Schwessinger et al., 2020), given that these contacts typically have a substantial impact over long-range regulation.

## 5.3 Conclusion

In conclusion, this study has investigated the efficiency of machine learning approaches in functional variant prediction across mammalian species. The proposed variant annotation pipeline, along with the machine learning approaches, can be extended to other livestock species of interest. These predictions from the model can in the future aid in prioritizing functional variants linked to traits of interest and for prioritizing variants in the prediction of Genomic Estimated Breeding Values (gEBVs). Consequently, these findings hold significant value for animal breeding and the improvement of animal welfare.

# References

Anaconda Software Distribution (4.10.3). (2021). [Computer software]. https://anaconda.com

Anderson, E., Peluso, S., Lettice, L. A., & Hill, R. E. (2012). Human limb abnormalities caused by disruption of hedgehog signaling. Trends in Genetics, 28(8), 364–373. https://doi.org/10.1016/j.tig.2012.03.012

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E. D., … Paten, B. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature, 587(7833), Article 7833. https://doi.org/10.1038/s41586-020-2871-y

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … National Eye Institute, N. (2015). A global reference for human genetic variation. Nature, 526(7571), Article 7571. https://doi.org/10.1038/nature15393

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18(10), Article 10. https://doi.org/10.1038/s41592-021-01252-x

Ayers, K. L., & Cordell, H. J. (2010). SNP Selection in Genome-Wide and Candidate Gene Studies via Penalized Logistic Regression. Genetic Epidemiology, 34(8), 879–891. https://doi.org/10.1002/gepi.20543

Baek, J., Lee, B., Kwon, S., & Yoon, S. (2018). LncRNAnet: Long non-coding RNA identification using deep learning. Bioinformatics, 34(22), 3889–3897. https://doi.org/10.1093/bioinformatics/bty418

Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. Nature Genetics, 40(3), Article 3. https://doi.org/10.1038/ng.78

Ben Braiek, M., Moreno-Romieux, C., Allain, C., Bardou, P., Bordes, A., Debat, F., Drögemüller, C., Plisson-Petit, F., Portes, D., Sarry, J., Tadi, N., Woloszyn, F., & Fabre, S. (2022). A Nonsense Variant in CCDC65 Gene Causes Respiratory Failure Associated with Increased Lamb Mortality in French Lacaune Dairy Sheep. Genes, 13(1), Article 1. https://doi.org/10.3390/genes13010045

Bertram, H., Wilhelmi, S., Rajavel, A., Boelhauve, M., Wittmann, M., Ramzan, F., Schmitt, A. O., & Gültas, M. (2023). Comparative Investigation of Coincident Single Nucleotide

Polymorphisms Underlying Avian Influenza Viruses in Chickens and Ducks. Biology, 12(7), Article 7. https://doi.org/10.3390/biology12070969

Billiard, S., Castric, V., & Llaurens, V. (2021). The integrative biology of genetic dominance. Biological Reviews of the Cambridge Philosophical Society, 96(6), 2925–2942. https://doi.org/10.1111/brv.12786

Bloom, K., Karunanidhi, A., Tobita, K., Hoppel, C., Thiels, E., Peet, E., Wang, Y., Basu, S., & Vockley, J. (2020). ACAD10 protein expression and Neurobehavioral assessment of Acad10-deficient mice. PLOS ONE, 15(12), e0242445. https://doi.org/10.1371/journal.pone.0242445

Boix, C. A., James, B. T., Park, Y. P., Meuleman, W., & Kellis, M. (2021). Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature, 590(7845), Article 7845. https://doi.org/10.1038/s41586-020-03145-z

Bourdon, C., Boussaha, M., Bardou, P., Sanchez, M.-P., Le Guillou, S., Tribout, T., Larroque, H., Boichard, D., Rupp, R., Le Provost, F., & Tosser-Klopp, G. (2021). In silico identification of variations in microRNAs with a potential impact on dairy traits using whole ruminant genome SNP datasets. Scientific Reports, 11(1), Article 1. https://doi.org/10.1038/s41598-021-98639-9

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H. H., & Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Research, 18(11), 1752–1762. https://doi.org/10.1101/gr.080663.108

Bouwman, A. C., Daetwyler, H. D., Chamberlain, A. J., Ponce, C. H., Sargolzaei, M., Schenkel, F. S., Sahana, G., Govignon-Gion, A., Boitard, S., Dolezal, M., Pausch, H., Brøndum, R. F., Bowman, P. J., Thomsen, B., Guldbrandtsen, B., Lund, M. S., Servin, B., Garrick, D. J., Reecy, J., … Hayes, B. J. (2018). Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nature Genetics, 50(3), Article 3. https://doi.org/10.1038/s41588-018-0056-5

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brody, T. (2016). Chapter 19—Biomarkers. In T. Brody (Ed.), Clinical Trials (Second Edition) (pp. 377–419). Academic Press. https://doi.org/10.1016/B978-0-12-804217-5.00019-9

Broekema, R. V., Bakker, O. B., & Jonkers, I. H. (2020). A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. Open Biology, 10(1), 190221. https://doi.org/10.1098/rsob.190221

Brown, A. A., Viñuela, A., Delaneau, O., Spector, T. D., Small, K. S., & Dermitzakis, E. T. (2017). Predicting causal variants affecting expression by using whole-genome

sequencing and RNA-seq from multiple human tissues. Nature Genetics, 49(12), Article 12. https://doi.org/10.1038/ng.3979

Brown, T. A. (2002). Mutation, Repair and Recombination. In Genomes. 2nd edition. Wiley-Liss. https://www.ncbi.nlm.nih.gov/books/NBK21114/

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current Protocols in Molecular Biology, 109(1), 21.29.1-21.29.9. https://doi.org/10.1002/0471142727.mb2129s109

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. PLoS Computational Biology, 8(12), e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Chamary, J., & Hurst, L. D. (2005). Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biology, 6(9), R75. https://doi.org/10.1186/gb-2005-6-9-r75

Chen, K. M., Wong, A. K., Troyanskaya, O. G., & Zhou, J. (2022). A sequence-based global map of regulatory activity for deciphering human genetics. Nature Genetics, 54(7), Article 7. https://doi.org/10.1038/s41588-022-01102-2

Chen, L., Fish, A. E., & Capra, J. A. (2018). Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. PLOS Computational Biology, 14(10), e1006484. https://doi.org/10.1371/journal.pcbi.1006484

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., Donnelly, S., … Karczewski, K. J. (2022). A genome-wide mutational constraint map quantified from variation in 76,156 human genomes (p. 2022.03.20.485034). bioRxiv. https://doi.org/10.1101/2022.03.20.485034

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, T., & Guestrin, C. (2016a). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., & Guestrin, C. (2016b). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, M., McCarl, B., & Fei, C. (2022). Climate Change and Livestock Production: A Literature Review. Atmosphere, 13(1), Article 1. https://doi.org/10.3390/atmos13010140

Cheruiyot, E. K., Haile-Mariam, M., Cocks, B. G., MacLeod, I. M., Xiang, R., & Pryce, J. E. (2021). New loci and neuronal pathways for resilience to heat stress in cattle. Scientific Reports, 11(1), Article 1. https://doi.org/10.1038/s41598-021-95816-8

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly, 6(2), 80–92. https://doi.org/10.4161/fly.19695

Collins, F. S., & Mansoura, M. K. (2001). The Human Genome Project. Cancer, 91(S1), 221–225. https://doi.org/10.1002/1097-0142(20010101)91:1+<221::AID-CNCR8>3.0.CO;2-9

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., … Flicek, P. (2022). Ensembl 2022. Nucleic Acids Research, 50(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10(2), giab008. https://doi.org/10.1093/gigascience/giab008

de Almeida, B. P., Reiter, F., Pagani, M., & Stark, A. (2022). DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nature Genetics, 54(5), Article 5. https://doi.org/10.1038/s41588-022-01048-5

De Gobbi, M., Viprakasit, V., Hughes, J. R., Fisher, C., Buckle, V. J., Ayyub, H., Gibbons, R. J., Vernimmen, D., Yoshinaga, Y., de Jong, P., Cheng, J.-F., Rubin, E. M., Wood, W. G., Bowden, D., & Higgs, D. R. (2006). A Regulatory SNP Causes a Human Genetic Disease by Creating a New Transcriptional Promoter. Science, 312(5777), 1215–1217. https://doi.org/10.1126/science.1126431

Dekker, J., & Mirny, L. (2016). The 3D genome as moderator of chromosomal communication. Cell, 164(6), 1110–1121. https://doi.org/10.1016/j.cell.2016.02.007

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), Article 4. https://doi.org/10.1038/nbt.3820

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. Frontiers of Computer Science, 14(2), 241–258. https://doi.org/10.1007/s11704-019-8208-z

Douville, C., Masica, D. L., Stenson, P. D., Cooper, D. N., Gygax, D. M., Kim, R., Ryan, M., & Karchin, R. (2016). Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). Human Mutation, 37(1), 28–35. https://doi.org/10.1002/humu.22911

Duchesne, A., Vaiman, A., Castille, J., Beauvallet, C., Gaignard, P., Floriot, S., Rodriguez, S., Vilotte, M., Boulanger, L., Passet, B., Albaric, O., Guillaume, F., Boukadiri, A., Richard, L., Bertaud, M., Timsit, E., Guatteo, R., Jaffrézic, F., Calvel, P., … Vilotte, J.-L. (2017). Bovine and murine models highlight novel roles for SLC25A46 in mitochondrial dynamics and metabolism, with implications for human and animal health. PLoS Genetics, 13(4), e1006597. https://doi.org/10.1371/journal.pgen.1006597

Dutta, P., Talenti, A., Young, R., Jayaraman, S., Callaby, R., Jadhav, S. K., Dhanikachalam, V., Manikandan, M., Biswa, B. B., Low, W. Y., Williams, J. L., Cook, E., Toye, P., Wall, E., Djikeng, A., Marshall, K., Archibald, A. L., Gokhale, S., Kumar, S., … Prendergast, J. G. D. (2020). Whole genome analysis of water buffalo and global cattle breeds highlights convergent signatures of domestication. Nature Communications, 11(1), Article 1. https://doi.org/10.1038/s41467-020-18550-1

Edinburgh Compute and Data Facility web site. (2021). [Computer software]. U of Edinburgh. <www.ecdf.ed.ac.uk>

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005a). The Sequence Ontology: A tool for the unification of genome annotations. Genome Biology, 6(5), R44. https://doi.org/10.1186/gb-2005-6-5-r44

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005b). The Sequence Ontology: A tool for the unification of genome annotations. Genome Biology, 6(5), R44. https://doi.org/10.1186/gb-2005-6-5-r44

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: Description, history and methods to detect structural variation. Briefings in Functional Genomics, 14(5), 305–314. https://doi.org/10.1093/bfgp/elv014

Esteller, M. (2011). Non-coding RNAs in human disease. Nature Reviews Genetics, 12(12), Article 12. https://doi.org/10.1038/nrg3074

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. Nature Biotechnology, 38(3), Article 3. https://doi.org/10.1038/s41587-020-0439-x

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. European Journal of Human Genetics, 24(8), Article 8. https://doi.org/10.1038/ejhg.2015.269

Flanigan, K. M., Dunn, D. M., von Niederhausern, A., Soltanzadeh, P., Howard, M. T., Sampson, J. B., Swoboda, K. J., Bromberg, M. B., Mendell, J. R., Taylor, L., Anderson, C. B., Pestronk, A., Florence, J., Connolly, A. M., Mathews, K. D., Wong, B., Finkel, R. S., Bonnemann, C. G., Day, J. W., … Weiss, R. B. (2011). Nonsense mutation-associated Becker muscular dystrophy: Interplay between exon definition and splicing regulatory elements within the DMD gene. Human Mutation, 32(3), 299–308. https://doi.org/10.1002/humu.21426

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., … Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology, 5(10), R80. https://doi.org/10.1186/gb-2004-5-10-r80

Ghaffari, M. H., Alaedin, M. T., Sadri, H., Hofs, I., Koch, C., & Sauerwein, H. (2021). Longitudinal changes in fatty acid metabolism and in the mitochondrial protein import system in overconditioned and normal conditioned cows: A transcriptional study using microfluidic quantitative PCR. Journal of Dairy Science, 104(9), 10338–10354. https://doi.org/10.3168/jds.2021-20237

Ghahramani, Z. (2004). Unsupervised Learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2—14, 2003, Tübingen, Germany, August 4—16, 2003, Revised Lectures (pp. 72–112). Springer. https://doi.org/10.1007/978-3-540-28650-9_5

Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E. M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., Burdett, T., Hayhurst, J., Baker, J., Ferrer, J., Gonzalez-Uriarte, A., Jupp, S., Karim, M. A., Koscielny, G., Machlitt-Northen, S., … Dunham, I. (2021). Open Targets Genetics: Systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Research, 49(D1), D1311–D1320. https://doi.org/10.1093/nar/gkaa840

Gibbs, E. M., Barthélémy, F., Douine, E. D., Hardiman, N., Shieh, P. B., Khanlou, N., Crosbie, R. H., Nelson, S. F., & Miceli, M. C. (2019). Large in-frame 5' deletions in DMD associated with mild Duchenne muscular dystrophy: Two case reports and a review of the literature. Neuromuscular Disorders : NMD, 29(11), 863–873. https://doi.org/10.1016/j.nmd.2019.09.009

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., Tam, P. K.-H., Tsui, L.-C., Waye, M. M. Y., Wong, J. T.-F., Zeng, C., Zhang, Q., Chee, M. S., Galver, L. M., Kruglyak, S., … Methods Group. (2003).

The International HapMap Project. Nature, 426(6968), Article 6968. https://doi.org/10.1038/nature02168

Glubb, D. M., Dholakia, N., & Innocenti, F. (2012). Liver expression quantitative trait loci: A foundation for pharmacogenomic research. Frontiers in Genetics, 3, 153. https://doi.org/10.3389/fgene.2012.00153

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R. S., Creixell, P., Karchin, R., Vazquez, M., Fink, J. L., Kassahn, K. S., Pearson, J. V., Bader, G., Boutros, P. C., Muthuswamy, L., Ouellette, B. F. F., Reimand, J., Linding, R., Shibata, T., Valencia, A., Butler, A., … Lopez-Bigas, N. (2013). Computational approaches to identify functional genetic variants in cancer genomes. Nature Methods, 10(8), 723–729. https://doi.org/10.1038/nmeth.2562

Grigoriadis, D., Perdikopanis, N., Georgakilas, G. K., & Hatzigeorgiou, A. G. (2022). DeepTSS: Multi-branch convolutional neural network for transcription start site identification from CAGE data. BMC Bioinformatics, 23(2), 395. https://doi.org/10.1186/s12859-022-04945-y

GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science (New York, N.Y.), 369(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776

Halldorsson, B. V., Eggertsson, H. P., Moore, K. H. S., Hauswedell, H., Eiriksson, O., Ulfarsson, M. O., Palsson, G., Hardarson, M. T., Oddsson, A., Jensson, B. O., Kristmundsdottir, S., Sigurpalsdottir, B. D., Stefansson, O. A., Beyter, D., Holley, G., Tragante, V., Gylfason, A., Olason, P. I., Zink, F., … Stefansson, K. (2022). The sequences of 150,119 genomes in the UK Biobank. Nature, 607(7920), Article 7920. https://doi.org/10.1038/s41586-022-04965-x

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., … Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research, 22(9), 1760–1774. https://doi.org/10.1101/gr.135350.111

Head, T., MechCoder, Louppe, G., Shcherbatyi, I., fcharras, Vinícius, Z., cmmalone, Schröder, C., nel215, Campos, N., Young, T., Cereda, S., Fan, T., rene-rex, Shi, K. (KJ), Schwabedal, J., carlosdanielcsantos, Hvass-Labs, Pak, M., … Fabisch, A. (2018). scikit-optimize/scikit-optimize: V0.5.2 [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.1207017

Health (US), N. I. of, & Study, B. S. C. (2007). Understanding Human Genetic Variation. In NIH Curriculum Supplement Series [Internet]. National Institutes of Health (US). https://www.ncbi.nlm.nih.gov/books/NBK20363/

Hebbar, P., & Sowmya, S. K. (2022). Genomic Variant Annotation: A Comprehensive Review of Tools and Techniques. In A. Abraham, N. Gandhi, T. Hanne, T.-P. Hong, T. Nogueira Rios, & W. Ding (Eds.), Intelligent Systems Design and Applications (pp. 1057–1067). Springer International Publishing. https://doi.org/10.1007/978-3-030-96308-8_98

Hickey, G., Paten, B., Earl, D., Zerbino, D., & Haussler, D. (2013). HAL: A hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics, 29(10), 1341–1342. https://doi.org/10.1093/bioinformatics/btt128

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., … Kent, W. J. (2006). The UCSC Genome Browser Database: Update 2006. Nucleic Acids Research, 34(Database issue), D590-598. https://doi.org/10.1093/nar/gkj144

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., & Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods, 9(5), Article 5. https://doi.org/10.1038/nmeth.1937

Huang, C., Shuai, R., Baokar, P., Chung, R., Rastogi, R., Kathail, P., & Ioannidis, N. (2023). Personal transcriptome variation is poorly explained by current genomic deep learning models (p. 2023.06.30.547100). bioRxiv. https://doi.org/10.1101/2023.06.30.547100

Hubisz, M. J., Pollard, K. S., & Siepel, A. (2011). PHAST and RPHAST: Phylogenetic analysis with space/time models. Briefings in Bioinformatics, 12(1), 41–51. https://doi.org/10.1093/bib/bbq072

Hutchinson, A., Asimit, J., & Wallace, C. (2020). Fine-mapping genetic associations. Human Molecular Genetics, 29(R1), R81–R88. https://doi.org/10.1093/hmg/ddaa148

Hutchinson, A., Watson, H., & Wallace, C. (2020). Improving the coverage of credible sets in Bayesian genetic fine-mapping. PLOS Computational Biology, 16(4), e1007829. https://doi.org/10.1371/journal.pcbi.1007829

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., & Farh, K. K.-H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell, 176(3), 535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015

Jia, H., Park, S.-J., & Nakai, K. (2021). A semi-supervised deep learning approach for predicting the functional effects of genomic non-coding variations. BMC Bioinformatics, 22(6), 128. https://doi.org/10.1186/s12859-021-03999-8

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., Essers, J., Mitrovic, M., Ning, K.,

Cleynen, I., Theatre, E., Spain, S. L., Raychaudhuri, S., Goyette, P., Wei, Z., … Cho, J. H. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature, 491(7422), 119–124. https://doi.org/10.1038/nature11582

Karnuta, J. M., & Scacheri, P. C. (2018). Enhancers: Bridging the gap between gene control and human disease. Human Molecular Genetics, 27(R2), R219–R227. https://doi.org/10.1093/hmg/ddy167

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D., & Kent, W. J. (2003). The UCSC Genome Browser Database. Nucleic Acids Research, 31(1), 51–54.

Karollus, A., Mauermeier, T., & Gagneur, J. (2023). Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. Genome Biology, 24(1), 56. https://doi.org/10.1186/s13059-023-02899-9

Käser, T. (2021). Swine as biomedical animal model for T-cell research—Success and potential for transmittable and non-transmittable human diseases. Molecular Immunology, 135, 95–115. https://doi.org/10.1016/j.molimm.2021.04.004

Kaukonen, M., Quintero, I. B., Mukarram, A. K., Hytönen, M. K., Holopainen, S., Wickström, K., Kyöstilä, K., Arumilli, M., Jalomäki, S., Daub, C. O., Kere, J., & Lohi, H. (2020). A putative silencer variant in a spontaneous canine model of retinitis pigmentosa. PLoS Genetics, 16(3), e1008659. https://doi.org/10.1371/journal.pgen.1008659

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Proceedings of the 31st International Conference on Neural Information Processing Systems, 3149–3157.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., & Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Research, 28(5), 739–750. https://doi.org/10.1101/gr.227819.117

Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research, 26(7), 990–999. https://doi.org/10.1101/gr.200535.115

Kerimov, N., Hayhurst, J. D., Peikova, K., Manning, J. R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M. P., Kuzmin, I., Trevanion, S. J., Burdett, T., Jupp, S., Parkinson, H., Papatheodorou, I., Yates, A. D., Zerbino, D. R., & Alasoo, K. (2021). A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nature Genetics, 53(9), Article 9. https://doi.org/10.1038/s41588-021-00924-w

Khazeeva, G., Sablauskas, K., van der Sanden, B., Steyaert, W., Kwint, M., Rots, D., Hinne, M., van Gerven, M., Yntema, H., Vissers, L., & Gilissen, C. (2022). DeNovoCNN:

A deep learning approach to de novo variant calling in next generation sequencing data. Nucleic Acids Research, 50(17), e97. https://doi.org/10.1093/nar/gkac511

Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., & Gottesman, M. M. (2007). A “silent” polymorphism in the MDR1 gene changes substrate specificity. Science (New York, N.Y.), 315(5811), 525–528. https://doi.org/10.1126/science.1135308

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), Article 3. https://doi.org/10.1038/ng.2892

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. Science (New York, N.Y.), 308(5720), 385–389. https://doi.org/10.1126/science.1109557

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., & Carninci, P. (2006). CAGE: Cap analysis of gene expression. Nature Methods, 3(3), Article 3. https://doi.org/10.1038/nmeth0306-211

Koellner, C. M., Mensink, K. A., & Highsmith, W. E. (2018). Chapter 5—Basic Concepts in Human Molecular Genetics. In W. B. Coleman & G. J. Tsongalis (Eds.), Molecular Pathology (Second Edition) (pp. 99–120). Academic Press. https://doi.org/10.1016/B978-0-12-802761-5.00005-5

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., & Gao, G. (2007). CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Research, 35(suppl_2), W345–W349. https://doi.org/10.1093/nar/gkm391

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. Bioinformatics, 28(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kotsiantis, S. B. (2013). Decision trees: A recent overview. Artificial Intelligence Review, 39(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Kramer, O. (2013). K-Nearest Neighbors. In O. Kramer (Ed.), Dimensionality Reduction with Unsupervised Nearest Neighbors (pp. 13–23). Springer. https://doi.org/10.1007/978-3-642-38652-7_2

Krogh, A. (2008). What are artificial neural networks? Nature Biotechnology, 26(2), Article 2. https://doi.org/10.1038/nbt1386

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. Briefings in Bioinformatics, 14(2), 144–161. https://doi.org/10.1093/bib/bbs038

Kumar, S., Harmanci, A., Vytheeswaran, J., & Gerstein, M. B. (2020). SVFX: A machine learning framework to quantify the pathogenicity of structural variants. Genome Biology, 21(1), 274. https://doi.org/10.1186/s13059-020-02178-x

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Research, 46(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Lappalainen, T., Li, Y. I., Ramachandran, S., & Gusev, A. (2024). Genetic and molecular architecture of complex traits. Cell, 187(5), 1059–1075. https://doi.org/10.1016/j.cell.2024.01.023

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology, 9(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Le, N. Q. K., Ho, Q.-T., Nguyen, V.-N., & Chang, J.-S. (2022). BERT-Promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection. Computational Biology and Chemistry, 99, 107732. https://doi.org/10.1016/j.compbiolchem.2022.107732

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nature Genetics, 47(8), Article 8. https://doi.org/10.1038/ng.3331

Li, C., Tian, D., Tang, B., Liu, X., Teng, X., Zhao, W., Zhang, Z., & Song, S. (2021). Genome Variation Map: A worldwide collection of genome variations across multiple species. Nucleic Acids Research, 49(D1), D1186–D1191. https://doi.org/10.1093/nar/gkaa1005

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009b). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., Zhao, T., Guan, D., Pan, Z., Bai, Z., Teng, J., Zhang, Z., Zheng, Z., Zeng, J., Zhou, H., Fang, L., & Cheng, H. (2023). Learning functional conservation between human and pig to decipher evolutionary mechanisms underlying gene expression and complex traits. Cell Genomics, 100390. https://doi.org/10.1016/j.xgen.2023.100390

Li, W., Yin, Y., Quan, X., & Zhang, H. (2019). Gene Expression Value Prediction Based on XGBoost Algorithm. Frontiers in Genetics, 10. https://www.frontiersin.org/articles/10.3389/fgene.2019.01077

Li, Y., Shi, W., & Wasserman, W. W. (2018). Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics, 19(1), 202. https://doi.org/10.1186/s12859-018-2187-1

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), Article 6. https://doi.org/10.1038/nrg3920

Liu, F., Li, H., Ren, C., Bo, X., & Shu, W. (2016). PEDLA: Predicting enhancers with a deep learning-based algorithmic framework. Scientific Reports, 6(1), Article 1. https://doi.org/10.1038/srep28517

Liu, K., Zhang, Y., Yu, Z., Xu, Q., Zheng, N., Zhao, S., Huang, G., & Wang, J. (2021). Ruminal microbiota–host interaction and its effect on nutrient metabolism. Animal Nutrition, 7(1), 49–55. https://doi.org/10.1016/j.aninu.2020.12.001

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., Zhan, X., 23andMe Research Team, HUNT All-In Psychiatry, Choquet, H., Docherty, A. R., Faul, J. D., Foerster, J. R., Fritsche, L. G., Gabrielsen, M. E., … Vrieze, S. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nature Genetics, 51(2), 237–244. https://doi.org/10.1038/s41588-018-0307-5

Liu, S., Gao, Y., Canela-Xandri, O., Wang, S., Yu, Y., Cai, W., Li, B., Xiang, R., Chamberlain, A. J., Pairo-Castineira, E., D'Mellow, K., Rawlik, K., Xia, C., Yao, Y., Navarro, P., Rocha, D., Li, X., Yan, Z., Li, C., … Fang, L. (2022). A multi-tissue atlas of regulatory variants in cattle. Nature Genetics, 54(9), Article 9. https://doi.org/10.1038/s41588-022-01153-5

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions (arXiv:1705.07874). arXiv. https://doi.org/10.48550/arXiv.1705.07874

Lunney, J. K., Van Goor, A., Walker, K. E., Hailstock, T., Franklin, J., & Dai, C. (2021). Importance of the pig as a human biomedical model. Science Translational Medicine, 13(621), eabd5758. https://doi.org/10.1126/scitranslmed.abd5758

Luo, R., Sedlazeck, F. J., Lam, T.-W., & Schatz, M. C. (2019). A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nature Communications, 10(1), Article 1. https://doi.org/10.1038/s41467-019-09025-z

Mahat, D. B., Kwak, H., Booth, G. T., Jonkers, I. H., Danko, C. G., Patel, R. K., Waters, C. T., Munson, K., Core, L. J., & Lis, J. T. (2016). Base-Pair Resolution Genome-Wide Mapping Of Active RNA polymerases using Precision Nuclear Run-On (PRO-seq). Nature Protocols, 11(8), 1455–1476. https://doi.org/10.1038/nprot.2016.086

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. Genome Biology, 20(1), 246. https://doi.org/10.1186/s13059-019-1828-7

Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M. M., Auton, A., Myers, S., Morris, A., Pirinen, M., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Hall, A. S., Hattersley, A. T., Hill, A. V. S., … Donnelly, P. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nature Genetics, 44(12), Article 12. https://doi.org/10.1038/ng.2435

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

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

Medina, I., De Maria, A., Bleda, M., Salavert, F., Alonso, R., Gonzalez, C. Y., & Dopazo, J. (2012). VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing. Nucleic Acids Research, 40(W1), W54–W58. https://doi.org/10.1093/nar/gks572

Meyerholz, D. K. (2016). Lessons learned from the cystic fibrosis pig. Theriogenology, 86(1), 427–432. https://doi.org/10.1016/j.theriogenology.2016.04.057

Miko, I. (2008). Genetic Dominance: Genotype-Phenotype Relationships | Learn Science at Scitable. http://www.nature.com/scitable/topicpage/genetic-dominance-genotype-phenotype-relationships-489

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. BMJ, 339, b2535. https://doi.org/10.1136/bmj.b2535

Morioka, M. S., Kawaji, H., Nishiyori-Sueki, H., Murata, M., Kojima-Ishiyama, M., Carninci, P., & Itoh, M. (2020). Cap Analysis of Gene Expression (CAGE): A Quantitative and Genome-Wide Assay of Transcription Start Sites. Methods in Molecular Biology (Clifton, N.J.), 2120, 277–301. https://doi.org/10.1007/978-1-0716-0327-7_20

Mostafavi, H., Spence, J. P., Naqvi, S., & Pritchard, J. K. (2022). Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery (p. 2022.05.07.491045). bioRxiv. https://doi.org/10.1101/2022.05.07.491045

Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. Human Molecular Genetics, 19(R2), R131–R136. https://doi.org/10.1093/hmg/ddq400

Mullen, M. P., Berry, D. P., Howard, D. J., Diskin, M. G., Lynch, C. O., Giblin, L., Kenny, D. A., Magee, D. A., Meade, K. G., & Waters, S. M. (2011). Single Nucleotide Polymorphisms in the Insulin-Like Growth Factor 1 (IGF-1) Gene are Associated with Performance in Holstein-Friesian Dairy Cattle. Frontiers in Genetics, 2, 3. https://doi.org/10.3389/fgene.2011.00003

Mulvey, B., Lagunas, T., & Dougherty, J. D. (2021). Massively Parallel Reporter Assays: Defining Functional Psychiatric Genetic Variants across Biological Contexts. Biological Psychiatry, 89(1), 76–89. https://doi.org/10.1016/j.biopsych.2020.06.011

Nassar, L. R., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, B. T., Lee, C. M., Muthuraman, P., Nguy, B., Pereira, T., Nejad, P., Perez, G., Raney, B. J., Schmelter, D., Speir, M. L., … Kent, W. J. (2023). The UCSC Genome Browser database: 2023 update. Nucleic Acids Research, 51(D1), D1188–D1195. https://doi.org/10.1093/nar/gkac1072

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7. https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021

Nelson, M. R., Marnellos, G., Kammerer, S., Hoyal, C. R., Shi, M. M., Cantor, C. R., & Braun, A. (2004). Large-Scale Validation of Single Nucleotide Polymorphisms in Gene Regions. Genome Research, 14(8), 1664–1668. https://doi.org/10.1101/gr.2421604

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 368(1620), 20120362. https://doi.org/10.1098/rstb.2012.0362

Noble, W. S. (2006). What is a support vector machine? Nature Biotechnology, 24(12), Article 12. https://doi.org/10.1038/nbt1206-1565

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., Lee, E. A., Zhu, Y., Plajzer-Frick, I., Pickle, C. S., Kato, M., Garvin, T. H., Pham, Q. T., Harrington, A. N., Akiyama, J. A., Afzal, V., Lopez-Rios, J., Dickel, D. E., Visel, A., & Pennacchio, L. A. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature, 554(7691), Article 7691. https://doi.org/10.1038/nature25461

Oubounyt, M., Louadi, Z., Tayara, H., & Chong, K. T. (2019). DeePromoter: Robust Promoter Predictor Using Deep Learning. Frontiers in Genetics, 10. https://www.frontiersin.org/articles/10.3389/fgene.2019.00286

Ouyang, J., Zhong, Y., Zhang, Y., Yang, L., Wu, P., Hou, X., Xiong, F., Li, X., Zhang, S., Gong, Z., He, Y., Tang, Y., Zhang, W., Xiang, B., Zhou, M., Ma, J., Li, Y., Li, G., Zeng, Z., … Xiong, W. (2022). Long non-coding RNAs are involved in alternative splicing and promote cancer progression. British Journal of Cancer, 126(8), Article 8. https://doi.org/10.1038/s41416-021-01600-w

Pabst, R. (2020). The pig as a model for immunology research. Cell and Tissue Research, 380(2), 287–304. https://doi.org/10.1007/s00441-020-03206-9

Park, P. J. (2009). ChIP–seq: Advantages and challenges of a maturing technology. Nature Reviews Genetics, 10(10), Article 10. https://doi.org/10.1038/nrg2641

Parsons, M. P., & Raymond, L. A. (2015). Chapter 20—Huntington Disease. In M. J. Zigmond, L. P. Rowland, & J. T. Coyle (Eds.), Neurobiology of Brain Disorders (pp. 303–320). Academic Press. https://doi.org/10.1016/B978-0-12-398270-4.00020-3

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825--2830.

Pinnapureddy, A. R., Stayner, C., McEwan, J., Baddeley, O., Forman, J., & Eccles, M. R. (2015). Large animal models of rare genetic disorders: Sheep as phenotypically relevant models of human genetic disease. Orphanet Journal of Rare Diseases, 10(1), 107. https://doi.org/10.1186/s13023-015-0327-5

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

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 36(10), Article 10. https://doi.org/10.1038/nbt.4235

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nature Reviews. Genetics, 11(7), 459–463. https://doi.org/10.1038/nrg2813

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). CatBoost: Unbiased boosting with categorical features. https://arxiv.org/abs/1706.09516v5

Qi, Y. (2012). Random Forest for Bioinformatics. In C. Zhang & Y. Ma (Eds.), Ensemble Machine Learning: Methods and Applications (pp. 307–323). Springer. https://doi.org/10.1007/978-1-4419-9326-7_11

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramachandran, A., Lumetta, S. S., Klee, E. W., & Chen, D. (2021). HELLO: Improved neural network architectures and methodologies for small variant calling. BMC Bioinformatics, 22(1), 404. https://doi.org/10.1186/s12859-021-04311-4

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016a). deepTools2: A next generation web server for deep-sequencing data analysis. Nucleic Acids Research, 44(W1), W160–W165. https://doi.org/10.1093/nar/gkw257

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016b). deepTools2: A next generation web server for deep-sequencing data analysis. Nucleic Acids Research, 44(W1), W160–W165. https://doi.org/10.1093/nar/gkw257

Rao, S., Yao, Y., & Bauer, D. E. (2021). Editing GWAS: Experimental approaches to dissect and exploit disease-associated genetic variation. Genome Medicine, 13(1), 41. https://doi.org/10.1186/s13073-021-00857-3

Rentzsch, P., Schubach, M., Shendure, J., & Kircher, M. (2021). CADD-Splice—Improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Medicine, 13(1), 31. https://doi.org/10.1186/s13073-021-00835-9

Richardson, N., Cook, I., Crane, N., Dunnington, D., & Francois, R. (2023). arrow: Integration to “Apache” “Arrow” [Computer software]. https://github.com/apache/arrow/

Ritchie, G. R. S., Dunham, I., Zeggini, E., & Flicek, P. (2014). Functional annotation of non-coding sequence variants. Nature Methods, 11(3), 294–296. https://doi.org/10.1038/nmeth.2832

Rojano, E., Seoane, P., Ranea, J. A. G., & Perkins, J. R. (2019). Regulatory variants: From detection to predicting impact. Briefings in Bioinformatics, 20(5), 1639–1654. https://doi.org/10.1093/bib/bby039

Rojas-Downing, M. M., Nejadhashemi, A. P., Harrigan, T., & Woznicki, S. A. (2017). Climate change and livestock: Impacts, adaptation, and mitigation. Climate Risk Management, 16, 145–163. https://doi.org/10.1016/j.crm.2017.02.001

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65, 386–408. https://doi.org/10.1037/h0042519

Salavati, M., Clark, R., Becker, D., Kühn, C., Plastow, G., Dupont, S., Moreira, G. C. M., Charlier, C., Clark, E. L., & on behalf of the BovReg consortium. (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), jkad108. https://doi.org/10.1093/g3journal/jkad108

Saunders, G. R. B., Wang, X., Chen, F., Jang, S.-K., Liu, M., Wang, C., Gao, S., Jiang, Y., Khunsriraksakul, C., Otto, J. M., Addison, C., Akiyama, M., Albert, C. M., Aliev, F., Alonso, A., Arnett, D. K., Ashley-Koch, A. E., Ashrani, A. A., Barnes, K. C., … Vrieze, S. (2022). Genetic diversity fuels gene discovery for tobacco and alcohol use. Nature, 612(7941), Article 7941. https://doi.org/10.1038/s41586-022-05477-4

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics, 19(8), Article 8. https://doi.org/10.1038/s41576-018-0016-z

Schipper, M., & Posthuma, D. (2022). Demystifying non-coding GWAS variants: An overview of computational tools and methods. Human Molecular Genetics, 31(R1), R73–R83. https://doi.org/10.1093/hmg/ddac198

Schubach, M., Re, M., Robinson, P. N., & Valentini, G. (2017). Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. Scientific Reports, 7(1), Article 1. https://doi.org/10.1038/s41598-017-03011-5

Schuster, S. L., & Hsieh, A. C. (2019). The untranslated regions of mRNAs in cancer. Trends in Cancer, 5(4), 245–262. https://doi.org/10.1016/j.trecan.2019.02.011

Schwessinger, R., Gosden, M., Downes, D., Brown, R. C., Oudelaar, A. M., Telenius, J., Teh, Y. W., Lunter, G., & Hughes, J. R. (2020). DeepC: Predicting 3D genome folding using megabase-scale transfer learning. Nature Methods, 17(11), Article 11. https://doi.org/10.1038/s41592-020-0960-3

Sehn, J. K. (2015). Chapter 9—Insertions and Deletions (Indels). In S. Kulkarni & J. Pfeifer (Eds.), Clinical Genomics (pp. 129–150). Academic Press. https://doi.org/10.1016/B978-0-12-404748-8.00009-5

Shabalina, S. A., & Spiridonov, N. A. (2004). The mammalian transcriptome and the function of non-coding DNA sequences. Genome Biology, 5(4), 105.

Sharma, A., Lee, J. S., Dang, C. G., Sudrajad, P., Kim, H. C., Yeon, S. H., Kang, H. S., & Lee, S.-H. (2015). Stories and Challenges of Genome Wide Association Studies in Livestock—A Review. Asian-Australasian Journal of Animal Sciences, 28(10), 1371–1379. https://doi.org/10.5713/ajas.14.0715

Sharma, J., Keeling, K. M., & Rowe, S. M. (2020). PHARMACOLOGICAL APPROACHES FOR TARGETING CYSTIC FIBROSIS NONSENSE MUTATIONS. European Journal of Medicinal Chemistry, 200, 112436. https://doi.org/10.1016/j.ejmech.2020.112436

Sharo, A. G., Hu, Z., Sunyaev, S. R., & Brenner, S. E. (2022). StrVCTVRE: A supervised learning method to predict the pathogenicity of human genome structural variants. The

American Journal of Human Genetics, 109(2), 195–209. https://doi.org/10.1016/j.ajhg.2021.12.007

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. Nucleic Acids Research, 29(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Shirley, M. D., Ma, Z., Pedersen, B. S., & Wheelan, S. J. (2015). Efficient “pythonic” access to FASTA files using pyfaidx [Preprint]. PeerJ PrePrints. https://doi.org/10.7287/peerj.preprints.970v1

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research, 15(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Slatkin, M. (2008). Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. Nature Reviews. Genetics, 9(6), 477–485. https://doi.org/10.1038/nrg2361

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). BioMart – biological queries made easy. BMC Genomics, 10(1), 22. https://doi.org/10.1186/1471-2164-10-22

Smith, T. S., Heger, A., & Sudbery, I. (2017). UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Research, gr.209601.116. https://doi.org/10.1101/gr.209601.116

Song, L., & Crawford, G. E. (2010). DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harbor Protocols, 2010(2), pdb.prot5384. https://doi.org/10.1101/pdb.prot5384

Stenson, P. D., Mort, M., Ball, E. V., Chapman, M., Evans, K., Azevedo, L., Hayden, M., Heywood, S., Millar, D. S., Phillips, A. D., & Cooper, D. N. (2020). The Human Gene Mutation Database (HGMD®): Optimizing its use in a clinical diagnostic or research setting. Human Genetics, 139(10), 1197–1207. https://doi.org/10.1007/s00439-020-02199-3

Steri, M., Idda, M. L., Whalen, M. B., & Orrù, V. (2018). Genetic Variants in mRNA Untranslated Regions. Wiley Interdisciplinary Reviews. RNA, 9(4), e1474. https://doi.org/10.1002/wrna.1474

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. Nature, 526(7571), 75–81. https://doi.org/10.1038/nature15394

Sullivan, P. F., Meadows, J. R. S., Gazal, S., Phan, B. N., Li, X., Genereux, D. P., Dong, M. X., Bianchi, M., Andrews, G., Sakthikumar, S., Nordin, J., Roy, A., Christmas, M. J., Marinescu, V. D., Wang, C., Wallerman, O., Xue, J., Yao, S., Sun, Q., … Lindblad-Toh, K. (2023). Leveraging base-pair mammalian constraint to understand genetic variation and human disease. Science, 380(6643), eabn2937. https://doi.org/10.1126/science.abn2937

Swanepoel, F., Stroebel, A., & Moyo, S. (Eds.). (2010). The Role of Livestock in Developing Communities: Enhancing Multifunctionality. SunBonani Media. https://doi.org/10.18820/9781928424819

Tait-Burkard, C., Doeschl-Wilson, A., McGrew, M. J., Archibald, A. L., Sang, H. M., Houston, R. D., Whitelaw, C. B., & Watson, M. (2018). Livestock 2.0 – genome editing for fitter, healthier, and more productive farmed animals. Genome Biology, 19(1), 204. https://doi.org/10.1186/s13059-018-1583-1

Talenti, A., & Prendergast, J. (2021). nf-LO: A Scalable, Containerized Workflow for Genome-to-Genome Lift Over. Genome Biology and Evolution, 13(9), evab183. https://doi.org/10.1093/gbe/evab183

Tayara, H., Tahir, M., & Chong, K. T. (2020). Identification of prokaryotic promoters and their strength by integrating heterogeneous features. Genomics, 112(2), 1396–1403. https://doi.org/10.1016/j.ygeno.2019.08.009

Teng, J., Gao, Y., Yin, H., Bai, Z., Liu, S., Zeng, H., Bai, L., Cai, Z., Zhao, B., Li, X., Xu, Z., Lin, Q., Pan, Z., Yang, W., Yu, X., Guan, D., Hou, Y., Keel, B. N., Rohrer, G. A., … Fang, L. (2024). A compendium of genetic regulatory effects across pig tissues. Nature Genetics, 56(1), 112–123. https://doi.org/10.1038/s41588-023-01585-7

The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G. R., Steering committee, Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. https://doi.org/10.1038/nature15393

Thiruvenkadan, A. K., Kandasamy, N., & Panneerselvam, S. (2008). Coat colour inheritance in horses. Livestock Science, 117(2), 109–129. https://doi.org/10.1016/j.livsci.2008.05.008

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., Sandstrom, R., Bates, D., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., … Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. Nature, 489(7414), 75–82. https://doi.org/10.1038/nature11232

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. Nature Reviews Methods Primers, 1(1), Article 1. https://doi.org/10.1038/s43586-021-00056-9

van Arensbergen, J., Pagie, L., FitzPatrick, V. D., de Haas, M., Baltissen, M. P., Comoglio, F., van der Weide, R. H., Teunissen, H., Võsa, U., Franke, L., de Wit, E., Vermeulen, M., Bussemaker, H. J., & van Steensel, B. (2019). High-throughput identification of human SNPs affecting regulatory element activity. Nature Genetics, 51(7), 1160–1169. https://doi.org/10.1038/s41588-019-0455-2

Virolainen, S. J., VonHandorf, A., Viel, K. C. M. F., Weirauch, M. T., & Kottyan, L. C. (2023). Gene–environment interactions and their impact on human health. Genes and Immunity, 24(1), 1–11. https://doi.org/10.1038/s41435-022-00192-6

Walker, B. N., & Biase, F. H. (2020). The blueprint of RNA storages relative to oocyte developmental competence in cattle (Bos taurus). Biology of Reproduction, 102(4), 784–794. https://doi.org/10.1093/biolre/ioaa015

Wall, J. D., & Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. Nature Reviews Genetics, 4(8), Article 8. https://doi.org/10.1038/nrg1123

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, 38(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, Q. S., Kelley, D. R., Ulirsch, J., Kanai, M., Sadhuka, S., Cui, R., Albors, C., Cheng, N., Okada, Y., Aguet, F., Ardlie, K. G., MacArthur, D. G., & Finucane, H. K. (2021). Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. Nature Communications, 12(1), Article 1. https://doi.org/10.1038/s41467-021-23134-8

Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A. P., van de Geijn, B., Reshef, Y., Márquez-Luna, C., O'Connor, L., Pirinen, M., Finucane, H. K., & Price, A. L. (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nature Genetics, 52(12), Article 12. https://doi.org/10.1038/s41588-020-00735-5

Witt, K. E., & Huerta-Sánchez, E. (2019). Convergent evolution in human and domesticate adaptation to high-altitude environments. Philosophical Transactions of the Royal Society B: Biological Sciences, 374(1777), 20180235. https://doi.org/10.1098/rstb.2018.0235

Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., & Fu, Y. (2019). Large Scale Incremental Learning. 374–382. https://openaccess.thecvf.com/content_CVPR_2019/html/Wu_Large_Scale_Incremental_Learning_CVPR_2019_paper.html

Xiang, R., MacLeod, I. M., Daetwyler, H. D., de Jong, G., O'Connor, E., Schrooten, C., Chamberlain, A. J., & Goddard, M. E. (2021). Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. Nature Communications, 12(1), Article 1. https://doi.org/10.1038/s41467-021-21001-0

Yang, J. (2014). Chapter Five—Enhanced Skeletal Muscle for Effective Glucose Homeostasis. In Y.-X. Tao (Ed.), Progress in Molecular Biology and Translational Science (Vol. 121, pp. 133–163). Academic Press. https://doi.org/10.1016/B978-0-12-800101-1.00005-3

Zeng, L., Cao, Y., Wu, Z., Huang, M., Zhang, G., Lei, C., & Zhao, Y. (2019). A Missense Mutation of the HSPB7 Gene Associated with Heat Tolerance in Chinese Indicine Cattle. Animals, 9(8), Article 8. https://doi.org/10.3390/ani9080554

Zeng, T., & Li, Y. I. (2022). Predicting RNA splicing from DNA sequence using Pangolin. Genome Biology, 23(1), 103. https://doi.org/10.1186/s13059-022-02664-4

Zhang, Z., Miteva, M. A., Wang, L., & Alexov, E. (2012). Analyzing Effects of Naturally Occurring Missense Mutations. Computational and Mathematical Methods in Medicine, 2012, 805827. https://doi.org/10.1155/2012/805827

Zheng, A., & Casari, A. (2018). Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly.

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., & Troyanskaya, O. G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature Genetics, 50(8), Article 8. https://doi.org/10.1038/s41588-018-0160-6

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 12(10), 931–934. https://doi.org/10.1038/nmeth.3547

Zhu, X., & Goldberg, A. B. (2009). Overview of Semi-Supervised Learning. In X. Zhu & A. B. Goldberg (Eds.), Introduction to Semi-Supervised Learning (pp. 9–19). Springer International Publishing. https://doi.org/10.1007/978-3-031-01548-9_2

Zook, J. M., & Salit, M. (2011). Genomes in a bottle: Creating standard reference materials for genomic variation - why, what and how? Genome Biology, 12(1), P31. https://doi.org/10.1186/1465-6906-12-S1-P31