The potential of regulatory variant prediction AI models to improve cattle traits

Zhao R, Owen R, Marr M, Siddharth J, Chue Hong N, Talenti A, Hassan MA, Prendergast JGD

Corresponding authors

RZ: Rongrong.Zhao@ed.ac.uk

JGDP: james.prendergast@roslin.ed.ac.uk

Abstract

Considerable progress has been made in developing machine learning models for predicting human functional variants, but progress in livestock species has been more limited. This is despite the disproportionate potential benefits such models could have to livestock research, from improving breeding values to prioritising functional variants at trait-associated loci. A key open question is what datasets and modelling approaches are most important to close this performance gap between species. In this work we have developed a new framework for predicting regulatory variants, that includes deriving key conservation metrics in cattle for the first time, variant annotation and model training. When trained on expression quantitative trait loci (eQTL) and massively parallel reporter assay (MPRA) data for human and cattle we show that this framework has a high performance at predicting regulatory variants, with a maximum area under the receiver operating characteristic curve (AUROC) score of 0.92 in human and 0.81 in cattle. We explore various approaches to further close this performance gap, including integrating advanced DNA sequence models and generating extra chromatin data, but illustrate that the best approach would be generating improved gold-standard sets of known cattle regulatory variants. Importantly we demonstrate both the human and cattle models substantially enrich for variants linked to important traits, with up to 25-fold enrichments for functional variants observed. Consequently, with our framework developed to be applicable across species these results not only demonstrate its potential utility for fine-

mapping functional variants and improving breeding values in cattle, but also its potential for wider use across animal species.

Introduction

Regulatory variants are crucial in shaping mammalian phenotypes and are estimated to account for approximately 60% of the genetic differences linked to human disease (Nica & Dermitzakis, 2013). However, while genome-wide association studies (GWAS) have successfully linked numerous areas of the genome to diseases and other traits, the causative genetic variants within those regions often remain unknown. This is primarily due to linkage disequilibrium (LD) among the variants, which confounds the identification of the genetic changes driving the trait. While intersecting GWAS results with the locations of known expression quantitative trait loci (eQTL) can indicate whether regulatory variation may underlie a GWAS peak, the precise causative variant still often remains unknown.

To address this issue there has been considerable interest in developing artificial intelligence (AI) models for predicting functional regulatory variants in the human genome. One of the earlier AI models for predicting potential functional regulatory variants was GWAVA (Ritchie et al., 2014), which used a random forest model to predict potential functional, non-coding variants from a range of genomic features, including chromatin datasets and sequence conservation scores. Depending on the background set of non-functional variants used in model training, the area under the receiver operating characteristic curve (AUROC) of these models, a combined measure of the model's performance across classification thresholds, ranged from 0.75 to 0.84. The following year DeepSEA, which used a convolutional neural network to predict chromatin datasets from local sequence alone (Zhou & Troyanskaya, 2015), was published. Using the predicted effects of sequence changes on chromatin profiles, the authors demonstrated it was possible to train models to predict functional variants. The reported AUROC of these models were up to 0.7, and higher than GWAVA when trained on the same datasets. More recently the original DeepSEA work was expanded to 21,907 chromatin profiles

across more than 1,300 cell lines and tissues. When this new model, Sei (K. M. Chen et al., 2022), was used to score variants by their predicted regulatory impact, known disease variants had predicted functional effect scores up to 6.5 times higher than background variants. In 2021 the Enformer model was released (Avsec et al., 2021), which like DeepSEA uses a convolutional neural network to predict omics data from sequence data. When the output of Enformer was used to train random forest classifiers to discriminate fine-mapped regulatory variants from background variants the observed AUROC was 0.75.

Consequently, significant progress has been made in predicting functional variants using machine learning approaches. However, most studies have focused on humans, where most omics data is available. Yet, understanding the functional consequences of variants in other species, particularly livestock, is also crucially important. The use of selective breeding in livestock means that the identification of functional variants could be leveraged to provide societal impacts more rapidly than in most other species. For example, previous work in cattle demonstrated that selecting variants based on functional-and-evolutionary trait heritability (FAETH) scores, which is defined by a range of regulatory and evolutionary features, had the potential to improve the rate of genetic gain for economically important phenotypes (Xiang et al., 2021). Importantly to improve genomic estimated breeding values, (gEBVs, the equivalent of polygenic risk scores in humans) one only needs to enrich for functional variants among the set of variants used to generate these values. Therefore, identification of functional variants in livestock species has the potential to rapidly lead to economic and public health gains.

A key constraint on developing machine learning models in livestock species is the lack of suitable training data. Studies like the genotype-tissue expression (GTEx) project (Lonsdale et al., 2013) have provided a comprehensive, gold-standard resource for studying gene expression, that have been refined to sets of fine-mapped regulatory variant using tools such as CaVEMaN (Brown et al., 2017). These high-quality datasets are highly suitable for machine learning model construction. With the aim of building a similar resource for livestock species, the farm animal genotype-tissue expression

(FarmGTEx) project initiated the exploration of regulatory variants in farm animals, such as pigs and cattle (Liu et al., 2022; Teng et al., 2024). The cattle GTEx project (Liu et al., 2022) analysed 7,180 publicly available RNA-Seq samples derived from 46 cattle breeds and breed combinations, covering 114 distinct tissues. The downstream eQTL study targeted 23 specific tissues, each comprising over 40 individuals. The farmGTEx dataset has proved invaluable for mapping eQTLs in livestock species but has challenges compared to the human GTEx project. Most notably it depended on combining disparate public datasets derived at different times in different breeds and stages, and relied on calling variants directly from the RNA-seq data, substantially reducing the number of assayed variants.

Another key challenge in applying these machine learning approaches to livestock species is the comparative scarcity of functional annotation tools and datasets for livestock species. Variant annotation constitutes a crucial step in genomic sequence analysis, serving to allocate functional information to DNA variants. Several variant annotation tools have been developed. For example, ANNOVAR (Wang et al., 2010) was developed based on the UCSC Genome Browser database and is specifically designed for annotating single nucleotide variants (SNVs) and insertions/deletions. Another widely used variant annotation tools based on Ensembl annotations is VEP, a robust toolset designed for annotating and prioritizing genomic variants in both coding and non-coding genomic regions (McLaren et al., 2016). While these annotation tools can offer abundant information about the variants, the annotations they provide tend to be more generic and may not be entirely customized for specific downstream analyses, such as machine learning approaches for predicting regulatory variants. Moreover, these annotation tools display a bias toward humans, underscoring the necessity for a tool capable of annotating other mammalian species. Likewise features, including key conservation metrics such as PhyloP and PhastCons scores (Pollard et al., 2010) that have been shown to be invaluable for identifying functional regulatory elements in humans (Sullivan et al., 2023), are not currently available for cattle and other livestock species.

The goal of this study was to assess the potential of developing machine learning models for predicting livestock functional variants. To do this we first developed a species-agnostic variant annotation pipeline tailored to annotating variants for downstream machine learning models. This included deriving key PhyloP and PhastCons conservation metrics for cattle for the first time. To establish a baseline for the performance of any particular modelling approach, we first applied it to human data then compared its performance in cattle. In this way we could quantify where the cattle models did or did not underperform, and therefore what novel datasets and approaches are likely most important to further improve cattle regulatory variant prediction. We investigated the effectiveness of different training and feature sets, and the inherent characteristics associated with a variant being predicted as regulatory. Finally, we highlight the substantial utility of both human and cattle models at predicting potential regulatory variants associated with specific downstream traits.

Results

Novel variant annotation pipeline and conservation metrics

To address the limitations of existing variant annotation tools for non-human species, we developed a species-agnostic and reusable variant annotation pipeline using Nextflow (Di Tommaso et al., 2017) (Figure 1). As shown in Table 1, the pipeline offers five categories of annotations, encompassing sequence conservation, variant position properties, VEP annotations, sequence context and predicted functional genomic scores based on the Enformer deep learning sequence-based model (Avsec et al., 2021).

Sequence conservation metrics have been shown to be particularly powerful annotations when trying to identify functional variation in the genome (Sullivan et al., 2023; Xiang et al., 2019). However, the major sequence conservation scores for livestock species are not available. To address this gap, we calculated the phastCons and phyloP conservation scores for cattle using the Cactus 241-way multiple alignment of vertebrates generated by the zoonomia consortium (Armstrong et al., 2020; Genereux et al., 2020). As shown in Supplementary Figure 1A, B, both of these newly-derived metrics are enriched in coding

regions, validating their ability to identify functional genomic regions. We have made these conservation metrics publicly available at Zenodo.

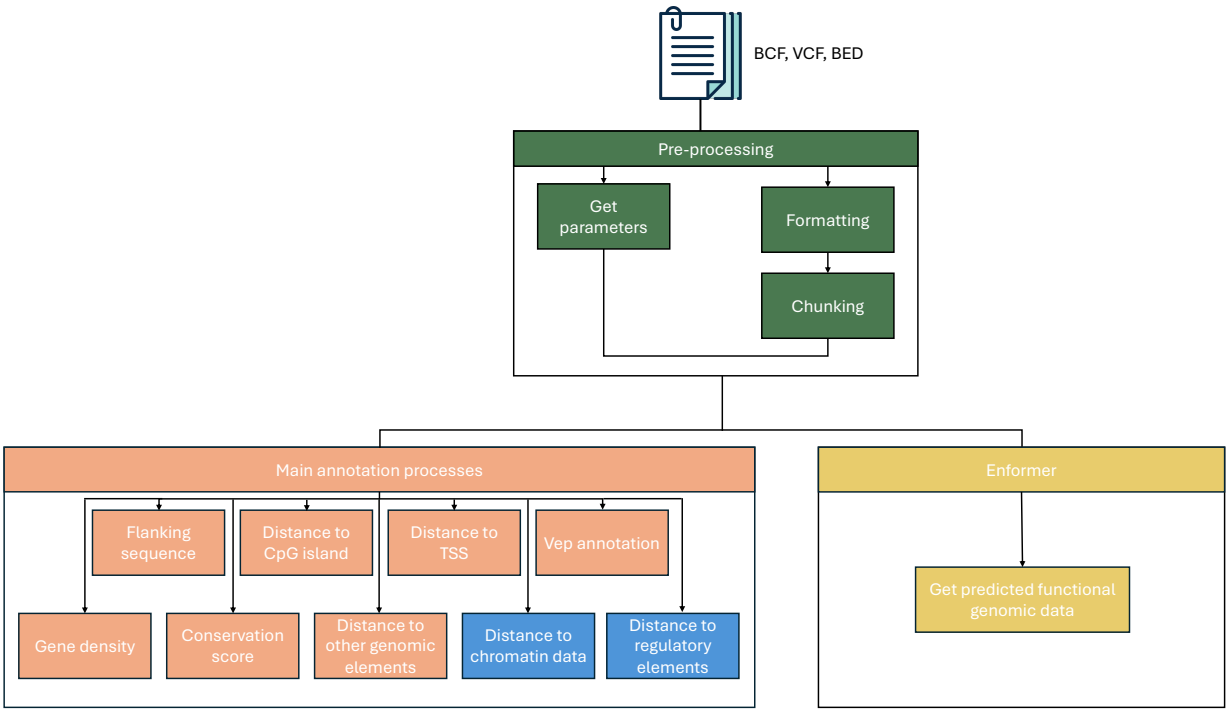


Figure 1. The workflow for the annotation pipeline. The workflow consists of three sections, pre-processing, the main annotation sub-workflow, and Enformer sub-workflow. The blue rectangles indicate annotations specific to humans.

Table 1. Variant annotation information

	Annotation	Data source	Main package	Number of features	Feature group
Sequence conservation	phastCons	Human: UCSC Cattle: Cactus, NCBI	Human: pyBigWig Other: Cactus, PhastCons	Human: 2, Cattle: 1	General features
	phyloP			Human: 2, Cattle: 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, cattle 3	
	Distance to regulatory features	Ensembl	ChIPpeakAnno	4 (Human only)	
	Chromosome	-	-	1	
	Variant position	-	-	1	
	Gene density (per Mb)	Ensembl	GenomicRanges	1	
VEP annotation	Consequence	Ensembl	VEP	1	
Sequence context	Allele change	-	-	1	
	5-mer flanking sequence	UCSC	Samtools	1	
Enformer	Predicted gene expression score	Enformer	TensorFlow	5313	Enformer features

Annotations highlighted in salmon colour indicate general features shared by both humans and cattle

The conservation of regulatory variant annotations across species

Next, we used our novel species-agnostic variation annotation pipeline to annotate known regulatory variants and matching background variants in both human and cattle. For humans, 590,778 known regulatory variants were obtained from CaVEMaN fine-mapping results of the GTEx data. For cattle 79,215 regulatory variants were obtained from cattle GTEx. To facilitate a direct comparison of annotations between regulatory and background variants, a matching number of background variants were randomly sampled for each species. As shown in Figure 2, the association of many features with regulatory variants were observed to be consistent across both species. For example, regulatory variants in both species often exhibit a distinctive sequence context compared to randomly selected background variants including a higher occurrence of G to A (C to T) allele change at the variant position (Figure 2A, B). Additionally, other positions within the 5-mer flanking sequences of regulatory variants demonstrated a tendency to harbor C:G base pairs (Figure 2I,J). Furthermore, these regulatory variants were more abundant in gene-dense regions and exhibited similar enrichments in proximity to specific genomic elements, such as CpG islands and transcription start sites (TSS) (Figure 2C to H). These observations underscore the common characteristics of regulatory variants across different species.

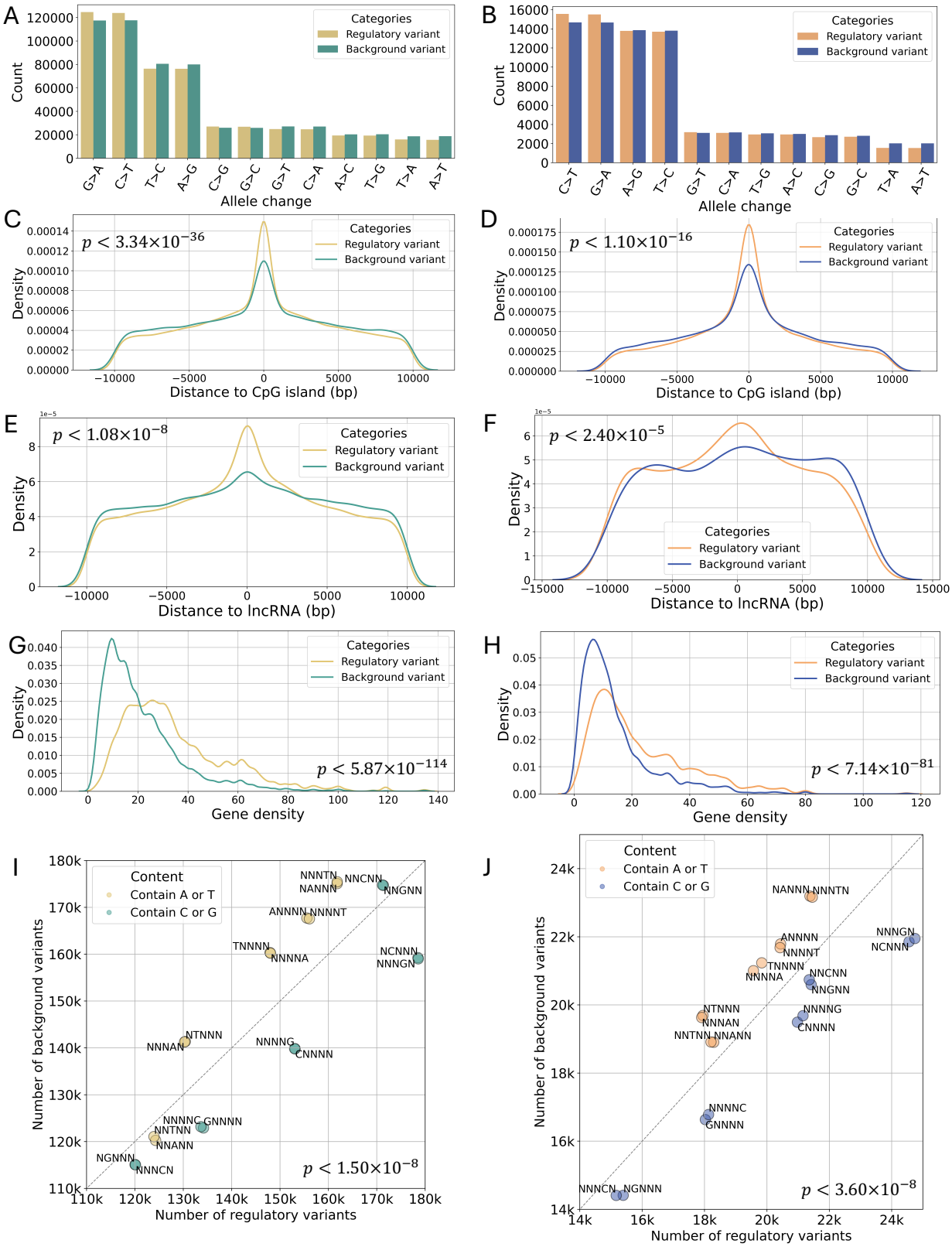

A
Count
Allele change
Categories
Regulatory variant
Background variant
G>A
C>T
T>C
A>G
C>G
G>C
G>T
C>A
A>C
T>G
T>A
A>T
B
Count
Allele change
Categories
Regulatory variant
Background variant
C>T
G>A
A>G
T>C
G>T
C>A
T>G
A>C
C>G
G>C
T>A
A>T
C
Density
Distance to CpG island (bp)
$p < 3.34\times10^{-36}$
Categories
Regulatory variant
Background variant
D
Density
Distance to CpG island (bp)
$p < 1.10\times10^{-16}$
Categories
Regulatory variant
Background variant
E
Density
Distance to lncRNA (bp)
$p < 1.08\times10^{-8}$
Categories
Regulatory variant
Background variant
F
Density
Distance to lncRNA (bp)
$p < 2.40\times10^{-5}$
Categories
Regulatory variant
Background variant
G
Density
Gene density
$p < 5.87\times10^{-114}$
Categories
Regulatory variant
Background variant
H
Density
Gene density
$p < 7.14\times10^{-81}$
Categories
Regulatory variant
Background variant
I
Number of background variants
Number of regulatory variants
Content
Contain A or T
Contain C or G
NNNTN
NANNN
NNCNN
NNGNN
ANNNN
NNNNT
TNNNN
NNNNA
NCNNN
NNNGN
NTNNN
NNNAN
NNNNG
CNNNN
NNNNC
GNNNN
NNTNN
NNANN
NGNNN
NNNCN
$p < 1.50\times10^{-8}$
J
Number of background variants
Number of regulatory variants
Content
Contain A or T
Contain C or G
NANNN
NNNTN
NNNGN
NCNNN
ANNNN
NNNNT
TNNNN
NNNNA
NNCNN
NNGNN
NTNNN
NNNAN
NNNNG
CNNNN
NNTNN
NNANN
NNNNC
GNNNN
NNNCN
NGNNN
$p < 3.60\times10^{-8}$

Figure 2 Common characteristics of human and cattle regulatory variants. Number of regulatory and background variants by their allele changes in (A) human and (B) cattle. Kernel density estimate (KDE) plots of distances from regulatory and background variants to CpG island within 10,000bp in (C) human and (D) cattle. KDE plots of distances from regulatory and background variants to lncRNA within 10,000bp in (E) human and (F) cattle. KDE plots of local gene density (genes/Mb) around regulatory and background variants in (G) human and (H) cattle. These annotations display significant differences between the groups in both human and cattle (Two-sample Kolmogorov-Smirnov test p-values are shown in the corresponding plots). Number of variants with different 5-mer flanking sequences between the regulatory and background data in (I) human and (J) cattle. 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 context. All types of 5-mer flanking sequences exhibit significant differences between the groups in both human and cattle (Chi-Squared test, p-values are shown in the plots).

The ability to predict regulatory variants in humans

To determine a baseline for using this annotation workflow and set of features to predict regulatory variants, we first applied it to a set of high confidence human regulatory variants. This human model effectively provides a glimpse into the potential future performance of cattle models if trained on the same datasets. The high-confidence human regulatory SNPs were those from the human GTEx project with a CaVEMaN causal probability exceeding 0.5 (45,987 variants) in any tissue. An equal number of matching background variants were selected from all variants tested in GTEx. These variants were annotated with general features as indicated in Table 1, excluding the Enformer features that are discussed in a later section. Different machine learning strategies, including Random Forest, CatBoost, XGBoost, LightGBM, GBDT and SVM, were explored and their performance is shown in Supplementary Figure 2. Among these experiments, the best performing high-confidence CatBoost model achieved an AUROC score of 0.89 at distinguishing fine-mapped regulatory variants from matching background variants.

Considering that the impact of regulatory variants is typically tissue-specific, we also conducted model training and testing across individual tissues. For this, fine-mapped human variants from five tissues relevant to cattle research (blood, liver, adipose, muscle, and breast) were retrieved from the human CaVEMaN dataset. Due to potential shortages in training data for these tissue-specific datasets, all CaVEMaN fine-mapped variants

were used, irrespective of causal probability. Subsequently, the background variants were down-sampled to match the number of regulatory variants retained in each tissue. As a result, the adipose, liver, blood, breast, and muscle sets comprised 64,378, 24,416, 48,774, 42,022, and 54,244 variants, respectively (each set includes both regulatory and background variants). As shown in Figure 3A, the human liver-specific model exhibited superior performance compared to the other tissue-specific models, achieving an AUROC score of 0.92. This performance also surpassed that of the model based on high-confidence cross-tissue eVariants (AUROC=0.89), despite a less stringent filter on causal probability being applied. The human liver is recognized for its relatively abundant tissue-specific eQTLs, attributed to several genes linked to liver-specific expression patterns (Glubb et al., 2012). Importantly, the liver is a relatively homogeneous tissue, which likely facilitates the precise delineation of regulatory variants compared to other tissues that are more complicated by cell type heterogeneity. Although direct comparisons to the performance of other models trained on different datasets should be done with caution, these results compare favourably to other published human models (Avsec et al., 2021; Ritchie et al., 2014; Zhou & Troyanskaya, 2015) and highlights the utility of our workflow for training models that substantially enrich for regulatory variants.

To understand which features impacted prediction outcomes, and therefore may be most relevant to training cattle models, a SHAP summary plot was generated for model interpretability. Figure 3B illustrates the top 30 most important features for the best-performing cross-tissue model, along with their corresponding impacts on the model's predictions. The most influential feature is the distance to the promoter. As shown in the plot, a moderate distance-to-promoter value (around 0, as distance includes both positive and negative values), indicating proximity to the promoter region, is linked to a higher probability of a variant being predicted as a regulatory variant. Other features that prominently contribute to the model outcomes include gene density, variant consequence, distance to TSS, flanking sequence, and conservation scores. In the tissue-specific models, in addition to features that exhibit importance across different models, such as gene density and distance to TSS, certain chromatin 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 (Supplementary Figure 3A). 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 (Supplementary Figure 3B). This suggests that tissue-specific chromatin data can improve tissue-specific model predictions.

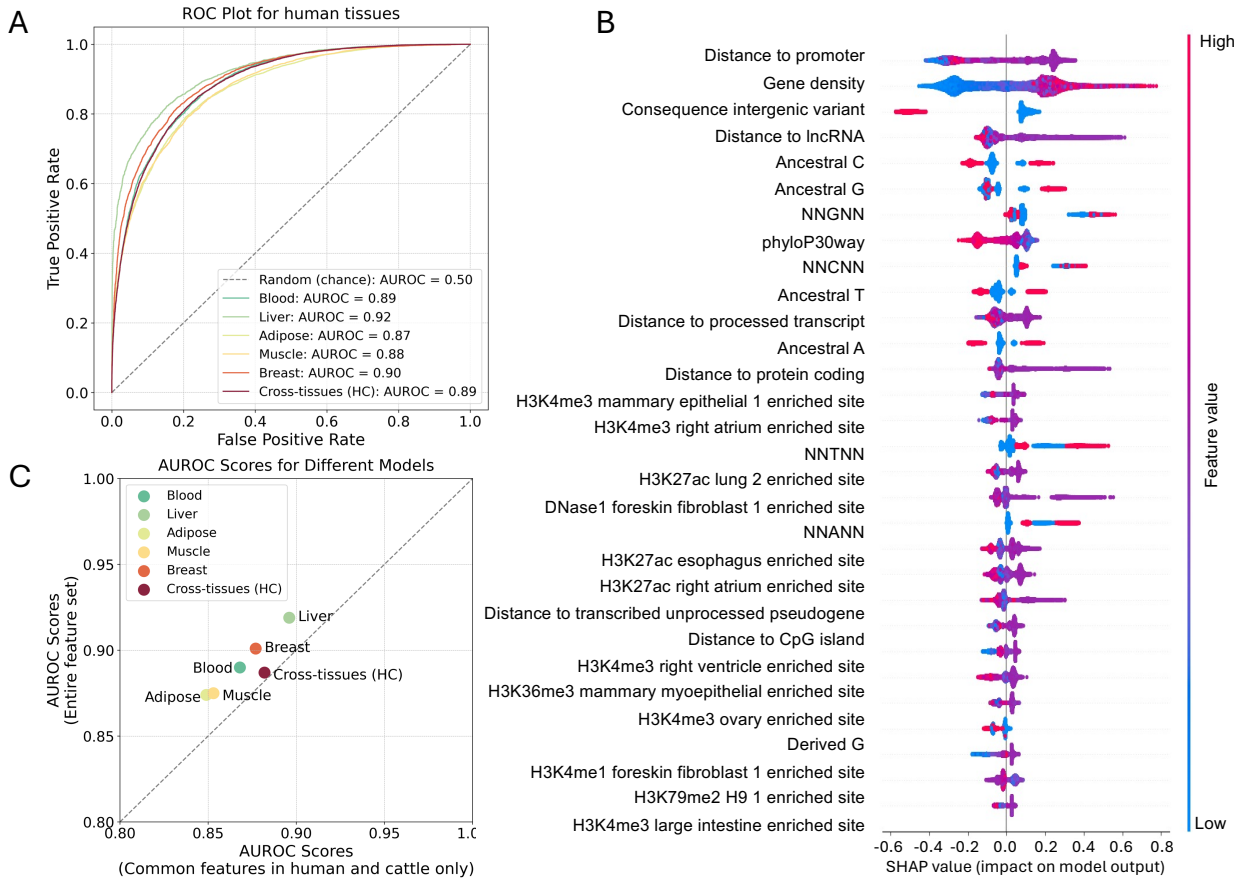


Figure 3 Human model performances. (A) Receiver Operating Characteristic (ROC) curves for human tissue-specific and cross-tissues high-confidence (HC) model. The numbers in the plot represent the area under the receiver operating characteristic (AUROC) scores for different models (ranging from 0 to 1), which evaluate the models' ability to distinguish regulatory variants. An AUROC closer to 1 indicates a better model. (B) The SHapley Additive exPlanations (SHAP) plot for the human high-confidence model. The Y axis of the SHAP plot displays the top 30 most important features in descending order based on their feature importance in the model. The colour scale shows the values of the features (high or low, as indicated by the colour bar) and the X axis their corresponding impact on the model's prediction (a positive value indicating the variant is more likely to be predicted as regulatory). (C) The comparison of the AUROC scores of human models trained on the entire set of general features, as indicated in Table 1, and human models

trained on the common features available for both human and cattle, as highlighted in salmon colour in Table 1. The grey dashed line represents parity. The X and Y axis are limited from 0.80 to 1.0 for better clarity of the scatter points.

Comparison of human and cattle baseline models

There is a comparative lack of chromatin and other omics datasets for most livestock species compared to humans, but notably, most of the important features in Figure 3B are readily available for non-human species. To examine this further we trained new human models, using the same set of human regulatory variants but excluding the human-specific omics datasets as features. This led to only modest drops in model performances (Figure 3C). While the human-specific features like distances to chromatin data and regulatory elements offer a slight boost to tissue-specific models (AUROC score increase of ~0.02), their impact on cross-tissues models is minimal (only ~0.005 improvement). Therefore, to directly compare the ability of predicting regulatory variants in different species, we trained cattle models on the same set of common features as highlighted by the salmon colour in Table 1. For cattle, 79,215 putative regulatory variants were obtained from the cattle GTEx project (see Method). The cross-tissue cattle model achieved an AUROC score of 0.72 (Figure 4A). Comparison of the feature importances for these human and cattle cross-tissue models trained on the common sets of over 90 features showed that the most important features determining the prediction results were largely consistent across the two models, with the top eight features being the same in both species (Figure 4B).

Tissue-specific regulatory variants for adipose, liver, blood, mammary, and muscle were also extracted for cattle. These sets comprised 6,308, 13,034, 16,328, 10,376, and 5,378 variants, respectively (each set includes both regulatory and background variants). Among cattle tissue-specific models, adipose and blood-specific models achieved the highest AUROC score of 0.73 (Figure 4A). In general, cattle models achieved poorer performance compared to human models when trained on the equivalent set of features. This suggests that the difference may in large part be driven by the difference in the quality of the regulatory variant sets used, not limitations in the sets of available features.

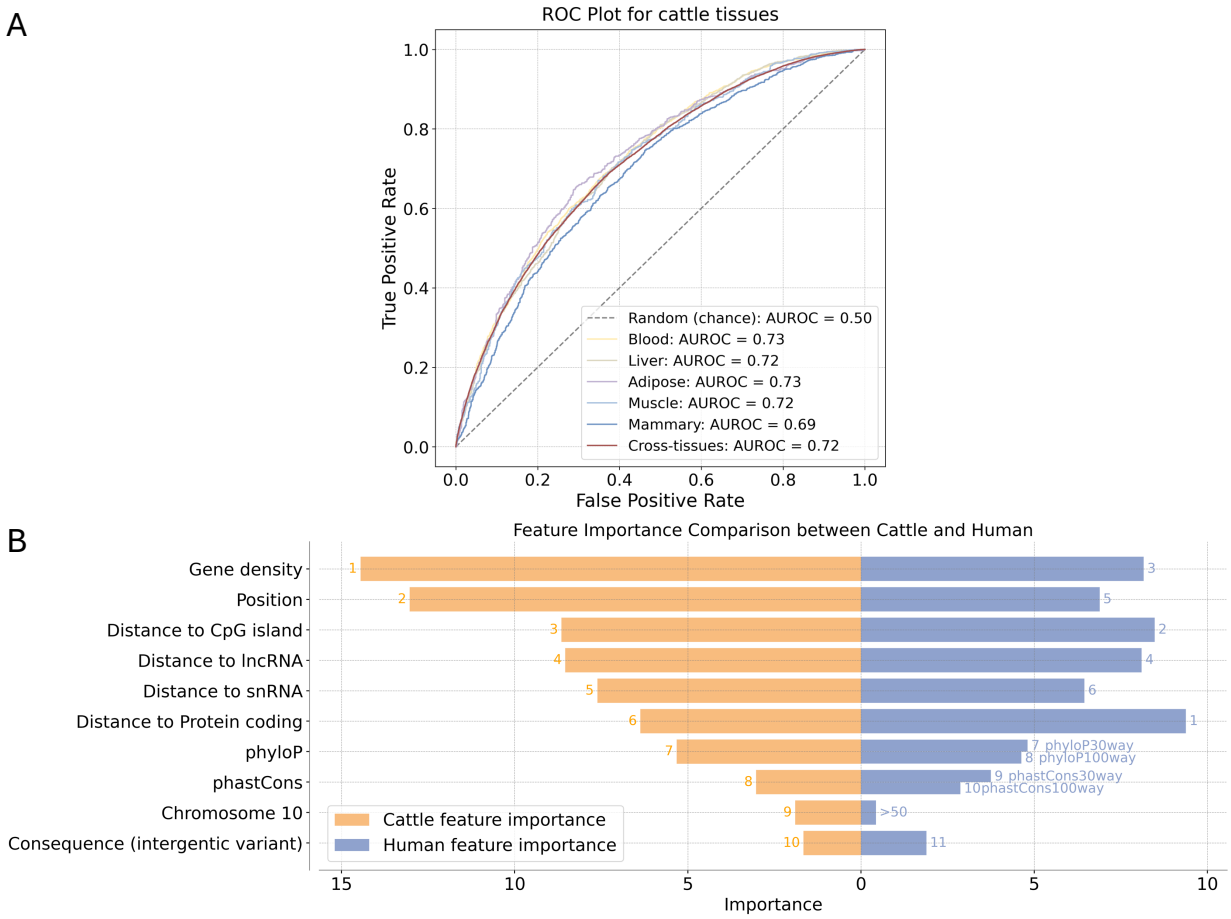


Figure 4 The comparison of cattle and human models trained on the common set of features. (A) ROC curves for cattle tissue-specific and cross-tissues models. The numbers in the plot legend represent the AUROC scores of different models. (B) Bar plot showing the top features in the cattle (orange) and human (blue) cross-tissues models trained on the common features. The X axis represents the relative feature importance in cattle and human models, respectively. The Y axis shows the top 10 features in the cattle model. The numbers next to the bars represent the rank of each feature. In humans, two phyloP conservation scores were available (phyloP30way and phyloP100way), ranking 7th and 8th, and two phastCons scores (phastCons30way and phastCons100way), ranking 9th and 10th. Cattle chromosome 10 shares synteny with human chromosomes 14 and 15 and these human chromosomes both rank lower than 50th in importance for the human model.

Human omics data does not improve cattle models

Previous studies have shown deep conservation of some regulatory elements in animals (Wong et al., 2020), potentially enabling annotation transfer between related species like

humans and cattle. To explore this, we characterized the ability of using human annotations to improve cattle regulatory variant prediction. Three approaches were tested: applying pre-trained human models to cattle variants, applying 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. These approaches achieved AUROC scores of 0.67, 0.69, and 0.71, respectively, which are lower than the cattle model trained on cattle annotations (AUROC = 0.72). Therefore, this outcome highlights that species-specific models outperform, even when fewer annotations are available. Human-based models, while capturing some general principles, may miss species-specific intricacies crucial for accurate predictions.

Enformer provides little benefit in predicting regulatory variants in either species

Recently a sequence-based model, Enformer, was developed. Enformer can predict a range of regulatory features from sequence alone, and the authors illustrated its potential at predicting regulatory variants by comparing regulatory feature predictions when variant alleles are changed in the input sequence (Avsec et al., 2021). Such models therefore provide a potential avenue to enhance feature sets from just genome sequences, and if it could, at least in part, be applied across species, it would have the potential to fill in gaps in omics data for less well characterised mammals. We consequently investigated the usefulness of features derived from the sequence-based Enformer model to inform both human and cattle predictions. The human high-confidence cross-tissues dataset was annotated with 5,313 Enformer features, i.e. the difference in predicted omics data between genomic sequences with reference and alternative alleles. We then trained CatBoost models using three feature sets: Enformer features only, general features only, and a combination of both. As shown in Figure 5A, the model trained with general features performed better than the one using only Enformer features. The combined model achieved comparable performance to the general features model, suggesting including Enformer features did not provide an improvement in model accuracy. We further applied the Enformer sub-workflow to annotate cattle tissue-specific variants. This aimed to explore whether these inferred chromatin features could improve cattle model predictions, which currently lack such data in their general annotations. However, incorporating these

additional Enformer features also did not improve the performance of the cattle tissue-specific models. The AUROC scores remained similar compared to the models trained using only general cattle features across all tissues, 0.72 in blood, 0.72 in liver, 0.73 in adipose, 0.72 in muscle, and 0.69 in mammary (Figure 5B). Our findings suggest that Enformer features offer limited improvement to regulatory variant prediction model performance in either species. These results are broadly consistent with recent work by Huang et al. highlighting limitations in some deep learning models' ability to capture the intricacies of individual genetic changes on gene expression regulation, especially those at distal sites (Huang et al., 2023).

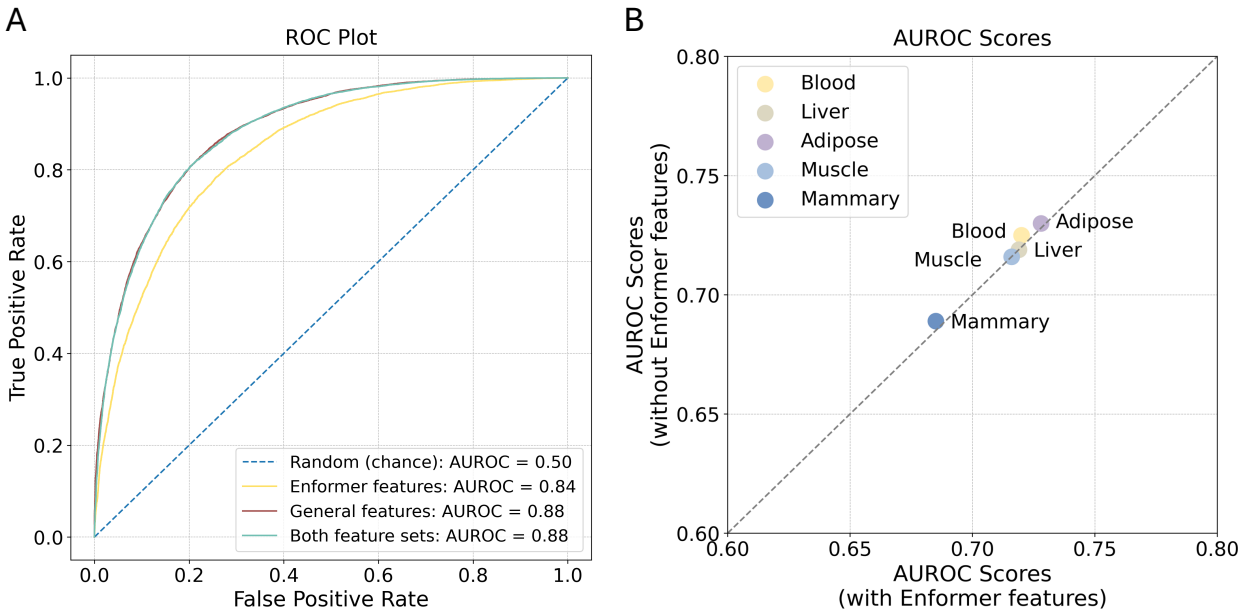


Figure 5 The comparison of human and cattle models trained with or without Enformer inferred features. (A) ROC curves for human high-confidence cross-tissues model trained using Enformer features, general features, and the combination of two feature sets. The numbers in the plot represent the AUROC scores of different models. (B) The comparison of the AUROC scores of cattle tissue-specific models trained with or without Enformer features. The grey dashed line represents parity. The X and Y axis are limited from 0.60 to 0.80 for better clarity of the scatter plot.

Prediction performance at distal regulatory elements

To assess our models' effectiveness at predicting regulatory variants across different genomic locations, we compared their performance on variants located within five distance ranges relative to TSSs across various tissues. These TSS bins were chosen to

ensure that each bin contained approximately the same number of variants within each tissue. As illustrated in Figure 6, the performance of all models, for both humans and cattle, declined as the distance of the variants from the TSS increased. This finding suggests that the models are more adept at capturing regulatory signals and making accurate predictions for variants closer to TSSs. An interesting exception is the human liver model, which consistently showed good performance across all distance ranges (Figure 6A), even for variants farther from the TSS, partly explaining its comparative higher performance. In other words, the human liver model performed better at predicting regulatory variants in distal regulatory elements such as enhancers. This might be due to the liver's relatively uniform cell composition compared to other tissues. The simpler cellular makeup might allow the model to identify cell-type specific enhancer-associated signals more readily.

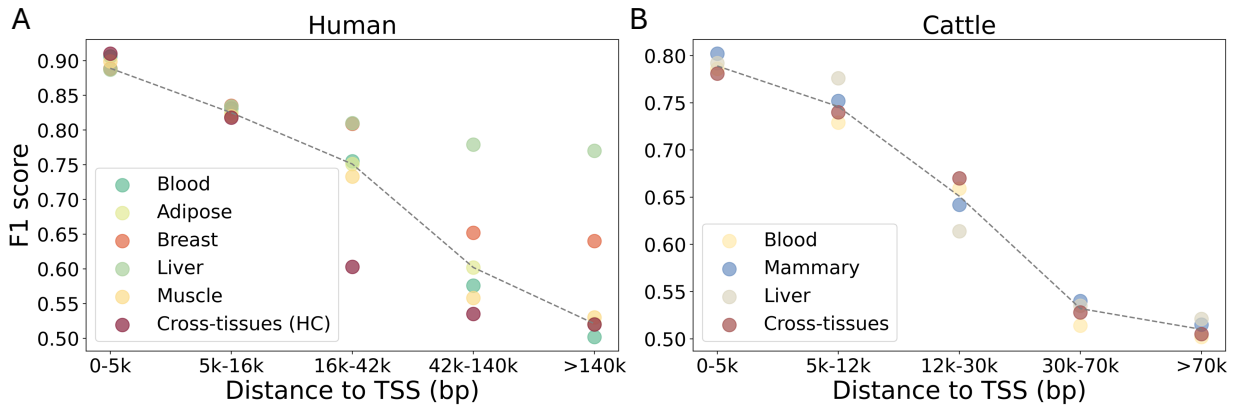


Figure 6 Human and cattle model performance using variants within different distance ranges to the TSS. (A) Human tissue-specific and cross-tissues high-confidence (HC) models performance based on variants in different TSS bins. The grey dashed line is plotted based on the median F1 score among the six datasets in each range. The F1 score representing a combined metric of the model's precision and recall. (B) Cattle tissue-specific and cross-tissues models performance based on variants in different TSS bins. The adipose and muscle datasets were excluded due to their relatively low variant counts. The grey dashed line is plotted based on the median F1 score among the four datasets in each range.

Model performance in predicting human and cattle MPRA regulatory variants

eQTLs have several limitations when used to train machine learning models. Most notably, only a subset of variants, such as those found above a certain allele frequency and within

a certain distance of genes (i.e. cis-eQTLs), are tested when characterising eQTL. Previous work has shown that the distribution of eQTLs in the genome is biased towards genes under less constraint and fewer functional annotations compared to variants linked to diseases and traits (Mostafavi et al., 2023). Precisely mapping the functional variant at a locus is also difficult, even when using statistical fine-mapping approaches, suggesting that eQTLs do not fully capture the spectrum of regulatory variants in the genome. Massively parallel reporter assays (MPRA) are an alternative approach for mapping regulatory variation in the genome that overcome some of these limitations. For example, approaches such as SuRE can test largely every heterozygote variant in the genome irrespective of its population-wise allele frequency or distance to known genes (van Arensbergen et al., 2019). Likewise, for most variants the confounder of LD is removed, as each variant is tested independently, allowing for the effect of individual variants on transcription to be characterised. MPRA approaches do though have their own limitations, most notably the fact that the variants are tested outside of their native sequence and chromatin environment. Likewise, they can only be carried out in suitably transfectable cell lines. These methods are therefore complementary approaches for defining regulatory variation.

In addition to using regulatory variants from the GTEx project, we also built the models using regulatory variants (raQTLs) based on the SuRE technique. For our human SuRE models, we included regulatory variants identified in two cell lines: K562 (a myelogenous leukemia cell line) and HepG2 (a liver hepatocellular carcinoma cell line) (van Arensbergen et al., 2019). We also trained a model on the combined dataset of these two lines. These three human SuRE models achieved AUROC scores of 0.76 (HepG2), 0.78 (K562), and 0.76 (combined dataset) on their respective test sets (Supplementary Figure 4).

For cattle modelling, we used a set of 66,501 SuRE raQTLs recently derived from cattle aortic endothelial cells (unpublished). To explore the potential optimal p-value threshold for defining raQTLs in cattle, we applied different FDR adjusted p-value thresholds. Equivalent numbers of background variants were randomly down-sampled from the

tested variants with adjusted p-values greater than 0.05. As shown in Figure 7, as the p-value threshold decreases, the AUROC score plateaus at around 0.81. However, noise becomes apparent when the p-value threshold is smaller than 1×10^{-10}, mainly due to the fewer variants included in the training and testing sets when below this cutoff.

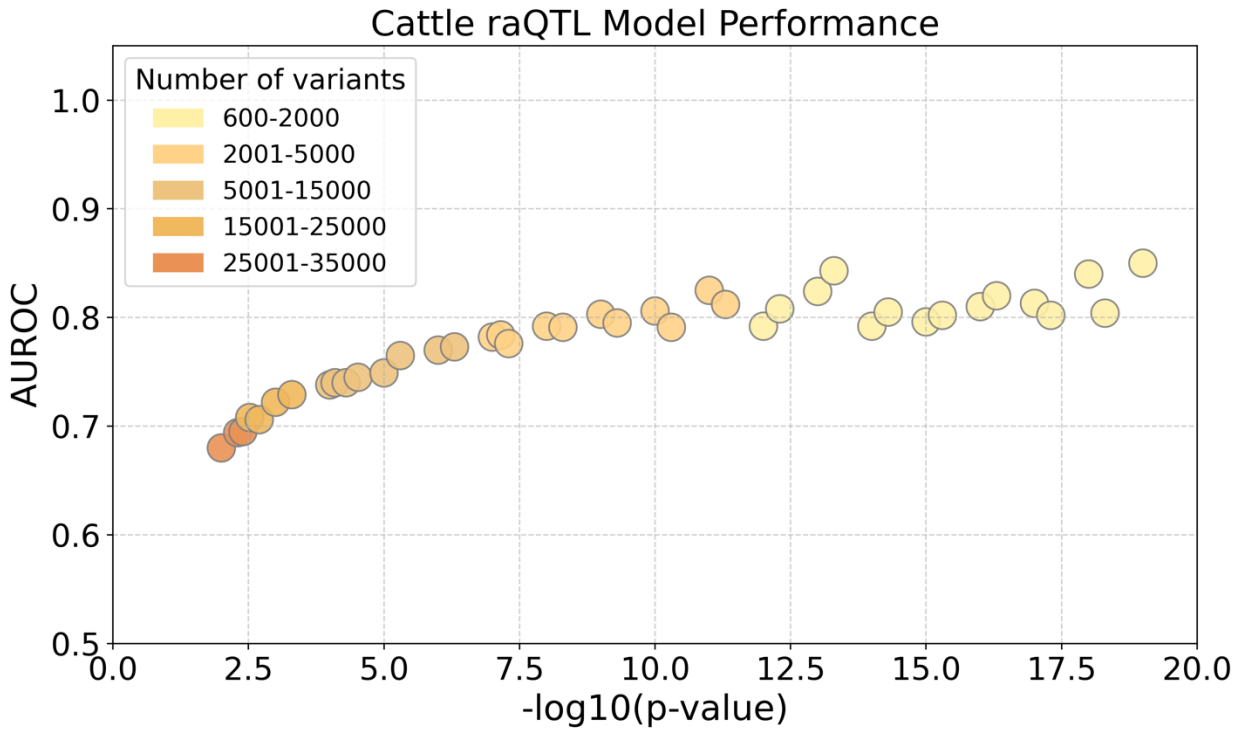


Figure 7 Cattle raQTL model performance when training and testing on regulatory variants selected based on different FDR adjusted p-value thresholds. For better clarity, the X axis is transformed to $-log_{10}^{p-value}$. The colour of each point represents the number of raQTL variants above the given threshold.

Our models substantially enrich for trait-associated variants in both species

A key motivation for training models that can predict regulatory variants is to identify which variants at loci linked to diseases and traits may be the functional drivers of the phenotypes. Likewise simply being able to enrich for functional variants among the set of genetic changes used to calculate genomic estimated breeding values is expected to improve their accuracy (Xiang et al., 2021). We therefore investigated the generalizability of our human and cattle regulatory models to predict functional GWAS variants. Of 28,918 human GWAS lead variants obtained from the credible sets analysis of the Open Target Genetics project (see Method), approximately 69% were identified as regulatory with a

prediction probability exceeding 0.5 using the cross-tissues high-confidence human GTEx model. This is an enrichment ratio of 2.38 compared to a background dataset of the same number of randomly selected variants, where only 29% were predicted as regulatory at the same threshold. The marked shift in prediction probability distributions between putative functional and background variants is shown in Figure 8A. To explore the impact of the prediction probability threshold for distinguishing regulatory variants in the model, different thresholds were utilised, and the corresponding ratio of the proportion of lead to background variants predicted as being regulatory was calculated. As shown in Figure 8B, it is possible to observe over 20-fold enrichments for functional variants at higher thresholds, suggesting these models can substantially enrich for variants driving downstream phenotypes.

To further assess the high-confidence human model's ability at predicting the regulatory effect of variants, we obtained 406 non-coding Mendelian disease-associated regulatory mutations (Smedley et al., 2016). After lifting these variants from GRCh37 to GRCh38, 345 single nucleotide variants on autosomes were retained for analysis. Notably, at a prediction probability cutoff of 0.5, 80.6% (278 out of 345) of these validated regulatory variants were predicted as regulatory, compared to only 19.7% (68 out of 345) of matching background variants randomly selected from the 1000 Genomes cohorts. Figure 8C illustrates the difference in the prediction probability distributions between these groups and Figure 8D explores how different prediction probability threshold impacts the ratios of regulatory variants predicted as regulatory to the background proportion. The high proportion of both GWAS and disease-associated variants predicted as regulatory by the model demonstrates the generalizability and effectiveness of our approach in prioritizing regulatory variants linked to downstream phenotypes.

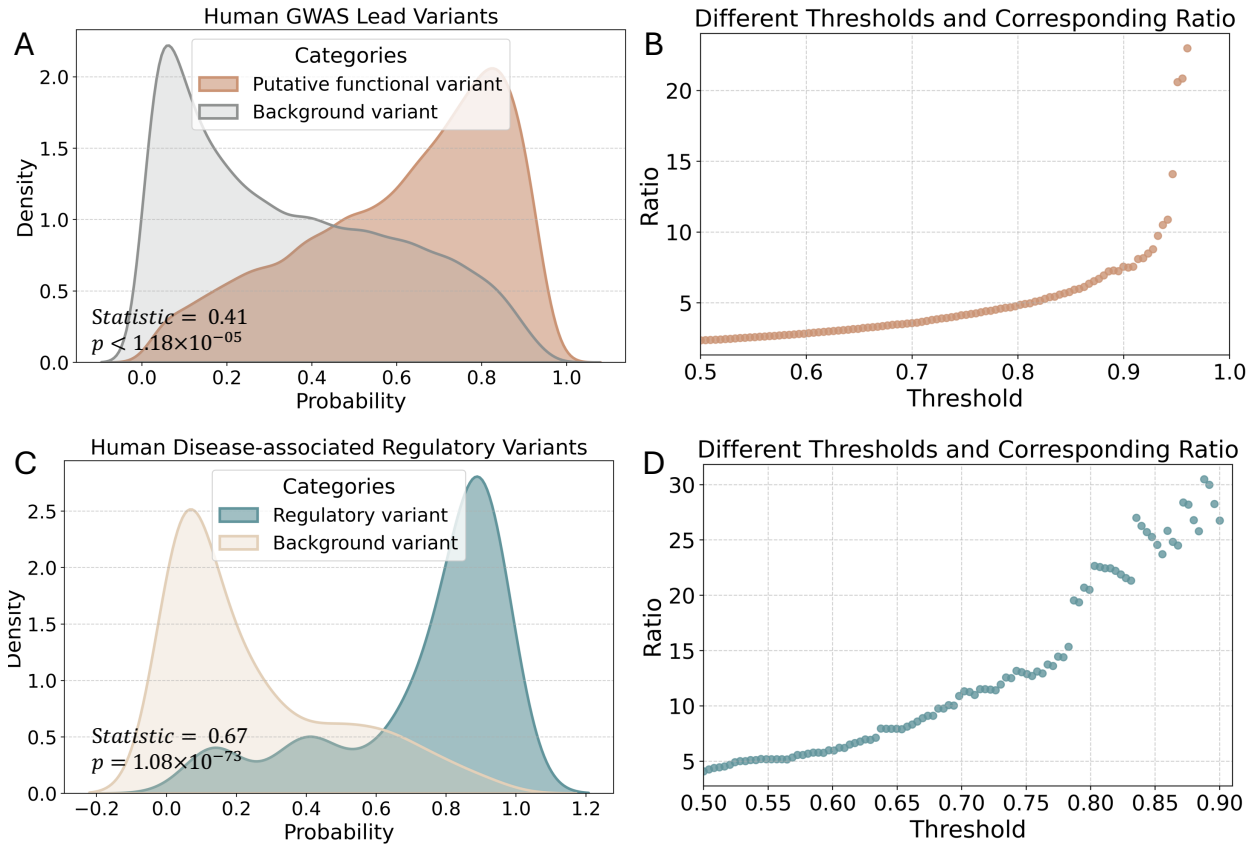


Figure 8 Prioritising human functional variants (A) Distributions of model derived prediction probabilities of putative human GWAS functional variants and background variants. The Kolmogorov-Smirnov statistic and p-value are shown in the plot. (B) Different prediction probability thresholds and their corresponding ratios of the proportion of GWAS lead variants predicted as regulatory to the proportion of background variants predicted as regulatory. (C) Distributions of model derived prediction probabilities of human disease-associated regulatory variants and background variants. The Kolmogorov-Smirnov statistic and p-value are shown in the plot. (D) Different prediction probability thresholds and their corresponding ratios of the proportion of human disease-associated regulatory variants predicted as regulatory to the proportion of background variants predicted as regulatory.

To explore the utility of our cattle models at prioritizing variants linked to important cattle phenotypes we obtained GWAS data for three cattle traits: body weight, stature, and body condition score (Reynolds et al., 2021). Using GCTA-COJO, we selected independently associated SNPs in each trait using the corresponding GWAS summary statistics data. Specifically, there were 344, 241, and 96 fine-mapped SNPs selected from the conditional analysis for body weight, stature, and body condition score, respectively. Table 2 summarizes the proportion of these variants predicted as regulatory using our different

cattle models. All models predicted a higher proportion of fine-mapped variants as regulatory compared to the equivalent number of matching background variants. At the default prediction probability threshold of 0.5, the ratios of the proportion of fine-mapped GWAS variants predicted as regulatory to the proportion of background variants predicted as regulatory ranged from 1.24 to 2.24. As shown in Table 2, the muscle-specific model achieved the best performance across all three traits among the tissue-specific models. Figure 9A,B,C illustrate the prediction probability distribution differences between fine-mapped and background data in the three traits. The cross-species model showed the most significant difference in prediction probability distributions between the fine-mapped and background variants. We further explored the impact of different prediction probability thresholds in the cross-species model on the ratios of the proportion of fine-mapped variants predicted as regulatory to background proportion. As shown in Figure 9D and E, 10-fold and 14-fold enrichment ratios are observed at higher thresholds in stature and body condition, respectively (results for body weight are shown in Supplementary Figure 5). This suggests that a higher prediction probability threshold could be useful for identifying high-confidence functional variants in GWAS datasets (at the cost of a higher false negative rate).

We identified variants that were consistently predicted as regulatory across different cattle models. Among the 96 body condition score variants, 18 of them were predicted as regulatory by all six models with a prediction probability exceeding 0.5. Similarly, 54 out of 344 body weight-associated SNPs and 36 out of 241 stature-associated SNPs were consistently predicted as regulatory by all six models with a prediction probability greater than 0.5. Supplementary figure 6 illustrates three examples. The chr14:23070083_C/T SNP, located near the TSS region of the *TGS1* gene, was predicted as a regulatory variant associated with stature (Supplementary figure 6A). This gene has been found to be associated with stature in both human and cattle populations (Zhao et al., 2015). The body weight-associated variant chr5:103942835_G/A is found in the intergenic region upstream of the *CD27* gene (Supplementary figure 6B), which has previously been associated with weight gain in mice (Englebert et al., 2024). Another predicted body condition score associated variant, chr2:125968341_G/A, was found near the *WDTC1*

gene (Supplementary figure 6C). This gene has been found associated with fat accumulation and obesity in humans (Ducos et al., 2017).

Table 2 The proportion of variants predicted as regulatory in three cattle traits using different models.

			Model type											
			Cross-tissues		Adipose		Blood		Liver		Mammary		Muscle	
Cattle Traits	Body condition score (96[a])	foreground	30 (31.3%)	2.14	34 (35.4%)	2.12	39 (40.6%)	2.05	37 (38.5%)	1.85	36 (37.5%)	2.12	38 (39.6%)	2.24
		background	14 (14.6%)		16 (16.7%)		19 (19.8%)		20 (20.8%)		17 (17.7%)		17 (17.7%)	
	Body weight (344[b])	foreground	109 (31.7%)	1.98	100 (29.1%)	1.67	90 (26.2%)	1.24	112 (32.6%)	1.58	123 (35.8%)	1.78	146 (42.4%)	2.03
		background	55 (16.0%)		60 (17.4%)		73 (21.2%)		71 (20.6%)		69 (20.1%)		72 (20.9%)	
	Stature (241[c])	foreground	81 (33.6%)	2.07	71 (29.5%)	1.87	78 (32.4%)	1.60	85 (35.3%)	1.67	84 (34.9%)	1.95	99 (41.1%)	1.98
		background	39 (16.2%)		38 (15.8%)		49 (20.3%)		51 (21.2%)		43 (17.8%)		50 (14.7%)	

[a,b,c] *Numbers of fine-mapped GWAS variants in body condition score, body weight, stature, respectively. Equivalent numbers of variants were selected as background variants in each trait for comparison.*

[*] *The numbers in the blue cells represent the count of variants predicted as regulatory in fine-mapped GWAS variants (foreground) using different models for each trait. The numbers in salmon cells are the count of variants predicted as regulatory in background variants using different models for each trait. The numbers in the purple cells represent the ratios of the proportion of foreground predicted as regulatory to background predicted as regulatory.*

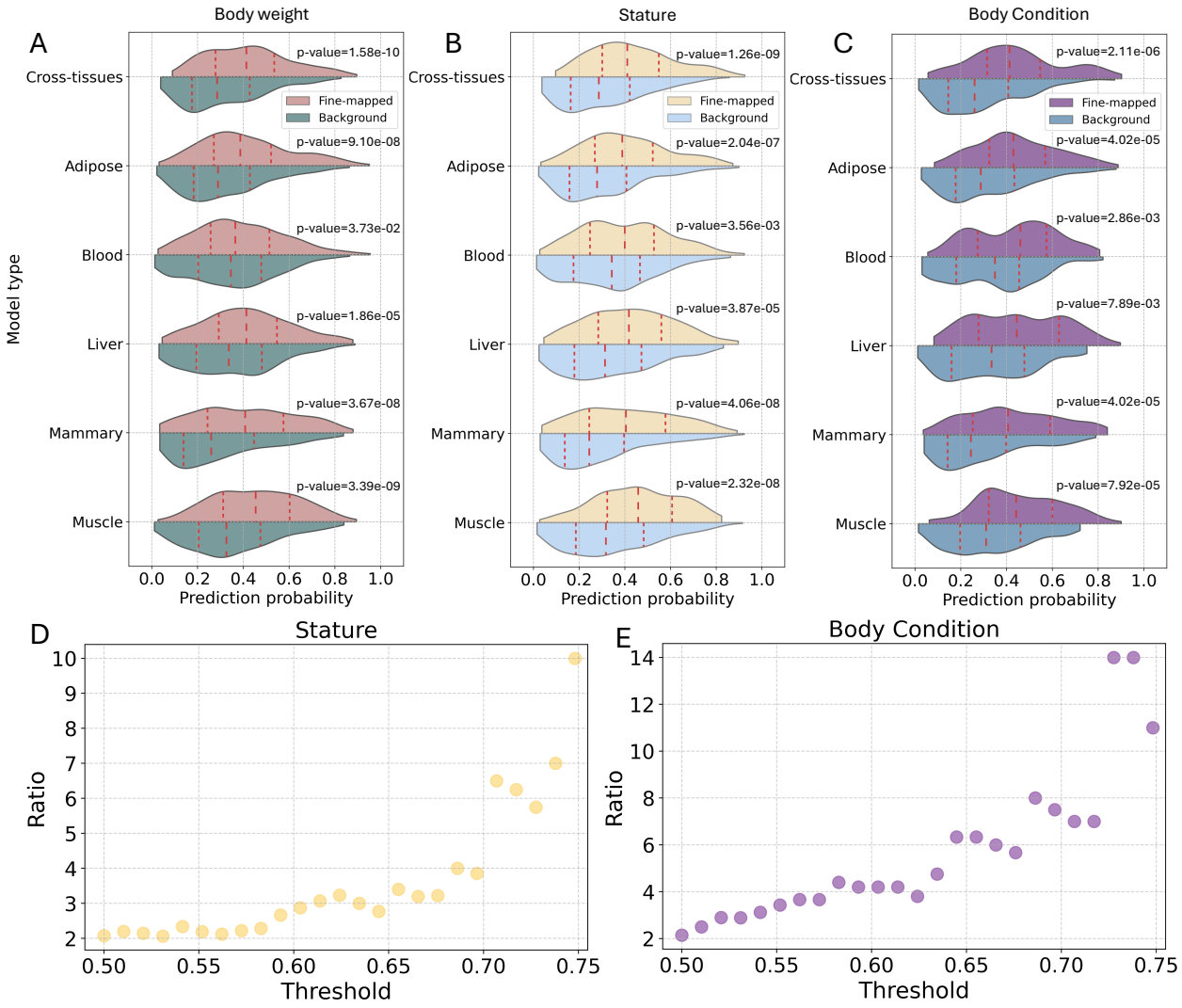


Figure 9 Prioritising cattle functional variants. Distributions of model derived prediction probabilities of cattle GWAS fine-mapped and background variants for (A) body weight, (B) stature, and (C) body condition score. The red dashed lines in the violin plots represent the median and quartile values. All prediction probabilities are significantly different between fine-mapped and background groups in each experiment. Their corresponding p-values are displayed in the figure (Two-sample Kolmogorov-Smirnov test). (D) Different prediction probability thresholds and their corresponding ratios of the proportion of cattle GWAS fine-mapped variants associated with stature predicted as regulatory to the proportion of background variants predicted as regulatory using the cross-species model. (E) Different prediction probability thresholds and their corresponding ratios of the proportion of cattle GWAS fine-mapped variants associated with body condition score predicted as regulatory to the proportion of background variants predicted as regulatory using the cross-species model.

Predicting regulatory elements linked to disease

These models effectively predict at which genomic locations a DNA change is expected to have a regulatory effect. Consequently, their utility may be broader than just predicting regulatory variants but also the location of key regulatory elements in the genome. One of the few examples in cattle where a distal regulatory element has been directly linked to a downstream phenotype, is an enhancer element downstream of the GC gene. In a previous study (Lee et al., 2021), disruption of this element by a copy number variant was linked to multiple traits including mastitis resistance, body condition score and milk production traits. Figure 10 A-C shows this locus and the blood model prediction probabilities for variants in this region. The variant with the highest prediction probability was chr6:86954490_C/T, that was not only the lead variant in the body condition score GWAS but also located at the enhancer region of the GC gene linked to the downstream phenotypes. So, despite the model not being trained using chromatin or GWAS data, it successfully identified the lead variant at this locus at the enhancer regulatory region. Together these results indicate the potential of our cattle models for prioritising regulatory variants and elements linked to important cattle traits.

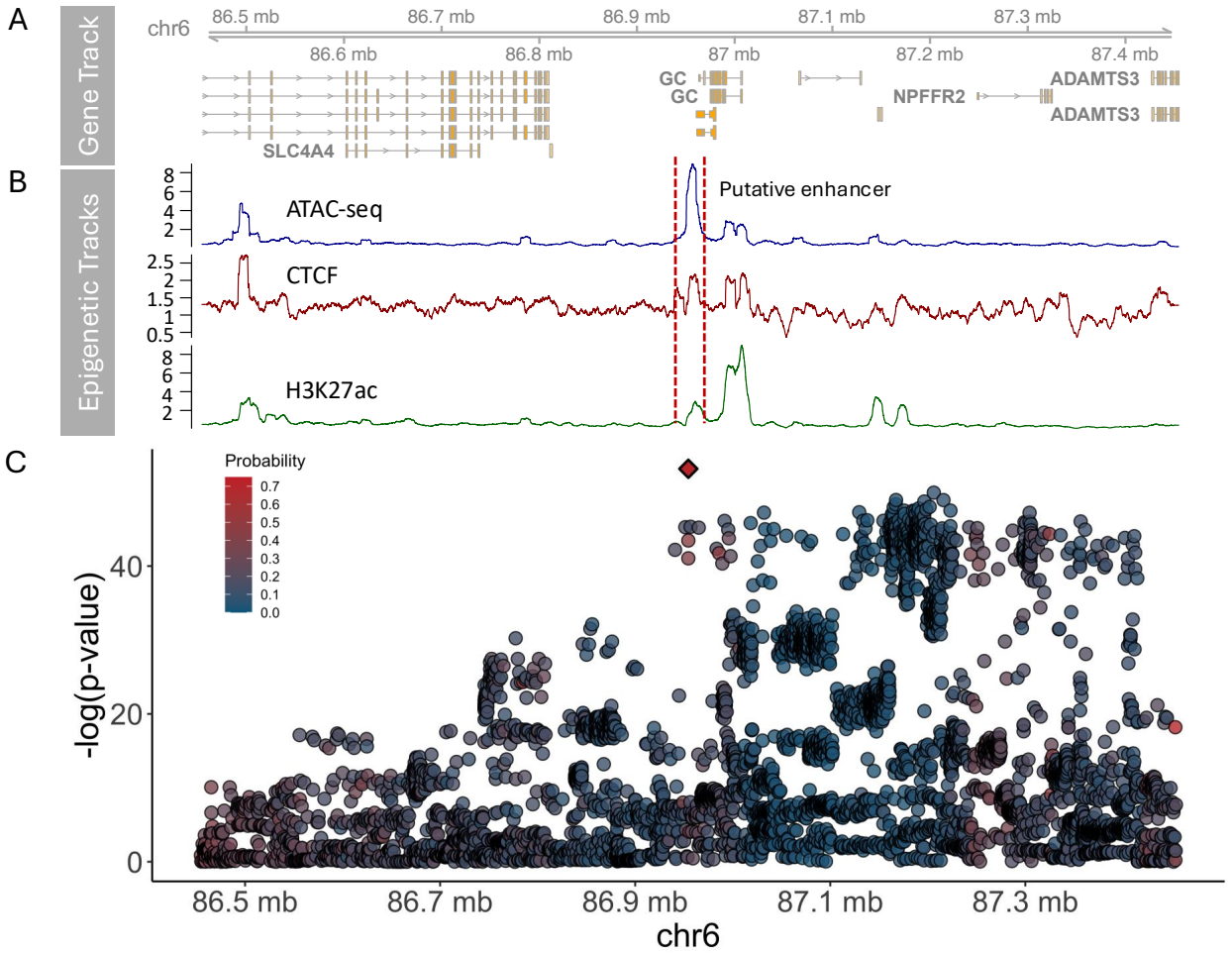


Figure 10 Regulatory predictions at the GC locus linked to traits including mastitis resistance and body condition score. (A) The genes and different isoforms at this locus. (B) ATAC-seq (Lee et al., 2021), CTCF and H3K27ac (Kern et al., 2021) tracks in this region, with the location of the putative trait-associated enhancer identified by Lee et al. indicated by the dashed lines. (C) Strength of association of cattle variants in the region with body condition score (Y axis), against their position (X axis). Colour indicates the prediction probabilities of each variant from the cattle blood model. Variant chr6:86954490_C/T is represented by a diamond.

Discussion

Identifying functional regulatory variants poses a significant challenge, largely attributed to their prevalence in comparatively poorer annotated non-coding regions and the presence of LD patterns among neighboring variants. This challenge is particularly acute in livestock species, where access to high-quality functional genomic data is relatively scarce. To address this challenge, we developed a variant annotation pipeline capable of

annotating variants with diverse features, ranging from sequence conservation to variant position properties, that is applicable across mammalian species. These annotations were then utilized as features to build human and cattle machine learning models for predicting variants with regulatory effects.

Models trained in both species were found to be effective at distinguishing regulatory variants from other variants in the genome, and notably similar features were observed to be important in both cases. Previous work to prioritise cattle functional variants had demonstrated the particular potential importance of conservation metrics for variant prioritization (Xiang et al., 2019). This previous study had though relied upon lifting over human conservation metrics to the cattle genome. This approach has potential issues, including that only around half of human bases can be directly mapped to a corresponding cattle base. In this study we generated cattle conservation metrics for the first time and illustrate how, consistent with this previous work, they were important features for predicting functional variants. Consequently, we believe these novel metrics will have wide-spread utility to cattle research.

Both human and cattle models performed better at predicting regulatory variants near the TSS than those located farther away. This can be attributed to the complexity of long-range regulatory interactions, and the fact that the sequence rules in the regions near TSSs are more pronounced, whereas at distal elements it 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 did consistently display good performance across different distance ranges to genes, even when testing on variants located far away from TSSs (>100kb). This proficiency in predicting regulatory variants not only within promoter regions but also at distal regulatory elements may be due to the lower rates of cellular heterogeneity in this tissue, reducing the confounding associated with looking across disparate cell types with their own regulatory profiles.

The application of our GTEx models to predict regulatory variants in humans and cattle demonstrates their effectiveness in prioritizing functional variants associated with traits. The enrichments of up to 14-fold for trait-associated variants highlights how these models could be particularly useful for aiding in the prioritization of variants for calculating genomic estimated breeding values (gEBVs) in livestock species. Unlike the direct identification of functional variants, prioritizing variants for gEBVs calculation does not necessarily require the precise identification of individual causal variants themselves, as long as those variants in LD with the causal ones are sufficiently well captured (Xiang et al., 2021). Hence, the utilization of these models in predicting gEBVs may not demand exceptionally high prediction accuracies. Simply enriching the dataset used to calculate gEBVs with functional variants, and tagging variants in high LD, is expected to improve their accuracies, leading to substantial cumulative benefits over time.

The analysis of the GC locus also highlights the potential of these models for fine-mapping functional regulatory elements at GWAS loci. Despite not being trained on the GWAS or chromatin data, the model identified the lead GWAS variant at the previously identified enhancer element as the variant with the highest regulatory potential. So, although the functional variant at this locus is thought to be a structural variant, this result highlights how relevant regulatory elements can potentially be prioritised using these models for further follow up study.

To date, deep learning-based models, especially those popular models in natural language processing (NLP), have been explored for understanding DNA sequences. However, the prediction results in our study illustrate that Enformer features did not provide additional discriminative information beyond our general features for both human and cattle models. This aligns with recent studies evaluating popular deep learning architectures, including Enformer, for predicting gene expression variation (Huang et al., 2023), as well as *in silico* investigations focusing on Enformer's ability to capture expression in promoters and enhancers (Karollus et al., 2023). These observations reported limitations in these models' performance. Specifically, they observed limited effectiveness in elucidating the impact of variants on gene expression and difficulties in

capturing the causal influence of distal enhancers on gene expression. Although these previous studies primarily focused on Enformer's ability to predict gene expression levels, rather than simply define the location of regulatory variants, these observations regarding Enformer's limitations align with the results of our study, highlighting the challenges deep learning models face in capturing the complex regulatory effects of individual base changes.

Notably the performance of models trained on the human SuRE data was lower than those trained on fine-mapped variants at eQTLs. SuRE data is though in theory less biased than that from eQTLs. Whereas cis-eQTL analyses are restricted to common variants close to genes, with the SuRE approach all polymorphic variants in the samples are tested. This comparative lack of bias may make distinguishing SuRE derived regulatory variants more challenging but may better reflect the true distribution of regulatory variants. A disadvantage of the SuRE approach is it is testing variants outwith their native genomic context, so that identified regulatory variants may not have impacts *in vivo*. This does though mean just the impact of sequence is assayed using this approach, removing other confounders, so that SuRE data holds the promise of potentially training better sequence-based models that capture the inherent impact of sequence effects.

The difference in performance between human and cattle models, even when using common features, suggests limitations in the cattle variant data. This is further supported by the minimal improvement (~0.02 AUROC score increase) observed in human models trained on all general features compared to models using only common human-cattle features. While incorporating chromatin data and other regulatory elements might hold promise, these findings suggest it may have a comparatively limited impact on improving cattle model performance. The core issue lies in the lack of a gold-standard regulatory variant set for training cattle models, which is largely responsible for their relatively low performance. While the cattle GTEx project serves as one of the most extensive reference resources for the cattle transcriptome available to date, it does have challenges. This includes a comparatively limited number of individuals and breeds for certain cell types,

the dependence on calling variants from the RNA-seq data, and difficulty of accurate variant imputation in underrepresented breeds. Additionally, unlike human GTEx, cattle GTEx data comes from diverse studies, potentially reducing its reliability due to variable batch effects. Together these factors may be leading to the difference in model performance between human and cattle. To an extent this is supported by the comparatively smaller gap between species in the performance of models trained on the more comparable SuRE datasets. Therefore, a comprehensive collection of high-quality cattle data, as well as data for other livestock species, encompassing high-quality functional variants would be indispensable. With the increased availability of high-quality training data expected to further improve the performance of livestock models in prioritizing novel functional variants, and ultimately improving advanced breeding approaches.

Methods

Variant annotation pipeline

We obtained a range of annotations for variants from different sources, including sequence conservation, variant position properties, VEP annotations, sequence context, and predicted functional genomic data based on Enformer. Bigwig files for four different human conservation scores (phastCons100way, phastCons30way, phyloP100way, phyloP30way) were downloaded from UCSC (Nassar et al., 2023). The Python package pyBigWig (Ramírez et al., 2016) was then used to extract the corresponding score for each variant from the bigwig files. For cattle, 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 cattle (Armstrong et al., 2020). The hierarchical alignment (HAL) format 241-way multiple alignment was first converted to multiple alignment format (MAF) by chromosome using the hal2maf tool from HAL toolkit (Hickey et al., 2013), with the cattle genome serving as the reference. Meanwhile, the general feature format (GFF) file for the corresponding cattle genome (Btau_5.0.1) included in the Cactus alignment was downloaded from NCBI (Sayers et al., 2022), and the coding sequences (CDS) were extracted from the GFF files. Then the CDS in bed format together with the MAF files for each chromosome were used as inputs to the

msa_view tool from the PHAST package (Hubisz et al., 2011) to generate four-fold degenerate sites (4d sites) in sufficient statistics (ss) format. The combined 4d sites file for the whole cattle genome was then randomly down-sampled to 40% of their original size to improve the computational efficiency in the downstream analysis. The down-sampled 4d sites and the phylogenetic tree of the HAL alignment were used as inputs to the phyloFit command (Hubisz et al., 2011) to estimate the neutral model. Subsequently, the conservation scores for cattle were computed using the phastCons and phyloP tools (Hubisz et al., 2011) with the neutral model serving as the phylogenetic model. Since the 241-way mammalian alignment employed the Btau_5.0.1 cattle genome assembly, we conducted a lift-over step to map the conservation scores to the updated BosTau9 assembly. Due to the lack of lift-over chain file between these two assemblies, nf-LO (Talenti & Prendergast, 2021) was used to generate the chain file. Subsequently, the USCS liftOver tool was used to perform the mapping.

For the variant position properties, distances from the variants to various genomic elements were calculated. The genomic locations of CpG islands for different species were downloaded from UCSC (Nassar et al., 2023). The 1554 different types of human chromatin data were obtained from Ensembl (Martin et al., 2023). Distances from the variants to the nearest CpG island and chromatin region (human only) were calculated using the bedtools closest function (Quinlan & Hall, 2010). The R library AnnotationHub was used to obtain transcription start sites (TSS) for different species, and only biotypes with greater than 1000 TSS were retained. The human regulatory features, including enhancers, transcription factors (TF), and CTCF binding sites, were obtained from Ensembl using the R library biomaRt (Smedley et al., 2009). Distances to the nearest transcription start site (TSS) and regulatory features (human only) were calculated using the ChIPpeakAnno R library (Zhu et al., 2010). Gene density per megabase pair (1Mb) for different species was calculated using the AnnotationHub R library (Gentleman et al., 2004) and GenomicRanges library (Lawrence et al., 2013). The VEP command line tool (McLaren et al., 2016) was used to obtain variant consequences. To capture the sequence context of the variants, except the allele change, we extracted the 5-mer flanking sequence centered on the target variant. Reference genomes for the two species in

FASTA format were obtained from UCSC (Nassar et al., 2023) and the Samtools faidx command was used to extract the flanking sequence from the reference genome (Li et al., 2009).

In addition to general annotations, 5313 predicted functional genomic data associated with gene expression and chromatin states for each variant were obtained using the Enformer model (Avsec et al., 2021). The pre-trained Enformer model was loaded from Tensorflow hub (https://tfhub.dev/deepmind/enformer/1), and the reference genome in FASTA format for the target species was downloaded from UCSC (Nassar et al., 2023). Python code for running the pre-trained Enformer model was adopted from the Enformer GitHub page (https://github.com/deepmind/deepmind-research/tree/master/enformer). The pre-trained model was employed to make predictions for the sequences (393,216 bp) centered on the reference and alternative alleles, respectively. Subsequently, the 5313-column scores for each variant were defined as the differences between the reference and alternative prediction results.

To facilitate the reusability of the aforementioned annotation approaches, a SNV annotation pipeline was developed using the Nextflow framework (Di Tommaso et al., 2017). These annotation approaches were encapsulated into separate processes, which can run in parallel to enhance annotation efficiency. They were organized into two sub-workflows, one for general annotations and another for Enformer scores. The default values of parameters and profiles for various execution environments were defined in the configuration files.

Genotype datasets

Various variant sets were used in the study to explore the utility of annotations in regulatory variant modelling. Human fine-mapped regulatory variants were obtained from the CaVEMaN dataset from the Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2020). We focused on autosomal SNPs, resulting in a set of 590,778 regulatory variants (foreground data) for analysis. To maintain consistency with the regulatory variants, an equivalent number of background variants were randomly down-

sampled from the GTEx reference data, limited to 1Mb around the TSS. For obtaining high-confidence regulatory variants, we filtered the 590,778 variants to those with a CaVEMaN causal probability > 0.5, resulting in 45,987 variants. Again, background variants were down-sampled to match with this size. Similarly, regulatory variants from cattle GTEx were acquired for cattle modelling. Lead variants, denoting those with the smallest nominal p-value for each gene across all tissues, were extracted. Furthermore, these variants were filtered based on their associated permutation p-values, using a threshold of 0.05. After filtering, 79,215 SNPs were retained as regulatory, 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.

For both humans and cattle, we obtained another set of regulatory variants (raQTL) based on the survey of regulatory elements (SuRE) technique. The human raQTLs were obtained from van Arensbergen et.al. (van Arensbergen et al., 2019). The original variants on hg19 were lifted over to hg38, resulting in 18,717 SNPs in the K562 cell line and 13,838 SNPs in the HepG2 cell line. The same number of background variants for these two cell lines were randomly sampled from the tested variants with SuRE signals greater than 4 and p-values greater than the threshold for defining raQTLs in the original paper (0.006192715 for K562 and 0.00173121 for HepG2). The cattle raQTLs were generated following the same approach as in van Arensbergen et.al. and is described in full detail in an accompanying paper. Instead of using the statistical approaches employed for identifying raQTLs in humans, our group used zero-inflated negative binomial regression models to analyze SuRE read count files. After filtering according to the adjusted p-value with a threshold of 0,05, there were 66,501 variants kept for further analysis. The same number of matching background variants were randomly sampled from the tested variants with adjusted p-value greater than 0.05.

To assess the effectiveness of human and cattle GTEx models in predicting regulatory effects of reported variants associated with traits, we used data from the Genome-wide association studies. For human GWAS data, we used the variants from the credible sets in Open Target Genetics (Ghoussaini et al., 2021). We downloaded the datasets from

Open Target Genetics and the Python package pyarrow (Richardson et al., 2023) was employed to convert the data into a Python dataframe from parquet format. Afterward, the variants were filtered based on a posterior probability threshold of 0.5, and those variants also present in the human high-confidence training set were removed, resulting in 28,918 variants. For comparison purposes, an equal number of background variants were randomly selected from the remaining variants with posterior probability smaller than 0.5. Another set of pathogenic regulatory variants associated with Mendelian diseases was obtained from Damina et al. (Smedley et al., 2016). The 406 single nucleotide variants were initially lifted over from the GRCh37 assembly to the GRCh38 assembly using the UCSC liftOver tool (Kuhn et al., 2013). After excluding those on sex chromosomes, 345 SNVs were retained for analysis. An equal number of background variants were randomly selected from the human 1000 genomes cohort (1000 Genomes Project Consortium et al., 2015).

For cattle GWAS data, we initially obtained the GWAS summary statistics data for three cattle traits: body condition score, body weight, and stature, from Edwardo et al (Reynolds et al., 2021). The originally reported variants on cattle genome UMD3.1 were lifted over to ARS-UCD1.2 using the mapping file provided in the paper. Subsequently we conducted conditional analysis using GCTA-COJO to select independently associated SNPs for each trait (Yang et al., 2012). To run GCTA-COJO, the initial summary statistics data were re-organized to match the input format required by the command. PLINK 2.0 (Chang et al., 2015) was used to generate the PLINK binary PED files, which served as an input for GCTA-COJO. To create the reference VCF files needed for input into PLINK 2.0, we first downloaded the reference VCF files from the 1000 Bull Genomes Project (Hayes & Daetwyler, 2019). Then these files were filtered using bcftools (Danecek et al., 2021) to include only a subset of samples that match the breeds included in the cattle GWAS data. After obtaining the associated SNPs for these traits, they were filtered based on their p-values using a threshold of 5e-8. Consequently, 96, 344 and 241 associated variants were retained for body condition score, body weight, and stature, respectively. For comparison, the same numbers of background SNPs with p-values greater than 0.05 were selected from the remaining summary statistics data.

Regulatory variant modelling

The variants were annotated using the variant annotation pipeline. These annotations first underwent feature engineering processes, performed using the Python scikit-learn package (Pedregosa et al., 2011). The conservation score annotations contained missing values, affecting approximately 0.9% of human variants and 0.8% of cattle variants. Given the small proportion, we opted to exclude variants with missing conservation scores from the dataset rather than employing imputation strategies. Subsequently, categorical features were encoded using different strategies. For sequence categorical features, such as 5-mer flanking sequence and allele change, we created a dictionary to map each nucleotide in the sequence to a binary string (A: 1000, C: 0100, G: 0010, T: 0001). The resulting string was then divided into separate integer columns, each indicating the presence or absence of a particular base. Categorical features without an ordinal relationship, such as chromosome and variant consequence, were encoded using one-hot encoding approach. After encoding, the transformed features were reintegrated into the feature table, replacing the original categorical features. This resulted in a 1680-column encoded general feature table for human variants and 96-column for cattle variant.

Seven machine learning algorithms were considered for model construction, Random Forest (Breiman, 2001), CatBoost (Prokhorenkova et al., 2017), XGBoost (T. Chen & Guestrin, 2016), LightGBM (Ke et al., 2017), Gradient Boosting Decision Trees (GBDT) (Friedman, 2001), and Support Vector Machines (SVM) (Cortes & Vapnik, 1995). Random Forest, SVM, and GBDT models were implemented using scikit-learn libraries, while CatBoost, XGBoost, and LightGBM utilized their respective Python packages. Each variant set, including both foreground and background variants, was randomly split into a training set (70%) and a test set (30%) using the train_test_split function from scikit-learn. The training set was used to train the models, and the separate test set was reserved for evaluating their performance on unseen data. To reduce bias and maximize the use of the training data, a 5-fold cross-validation approach was employed during model training. To further enhance model performance, hyper-parameter tuning was conducted using a combination of approaches: BayesSearchCV from the Python scikit-optimize library (Head et al., 2018) for automated optimization, and manual tuning for specific

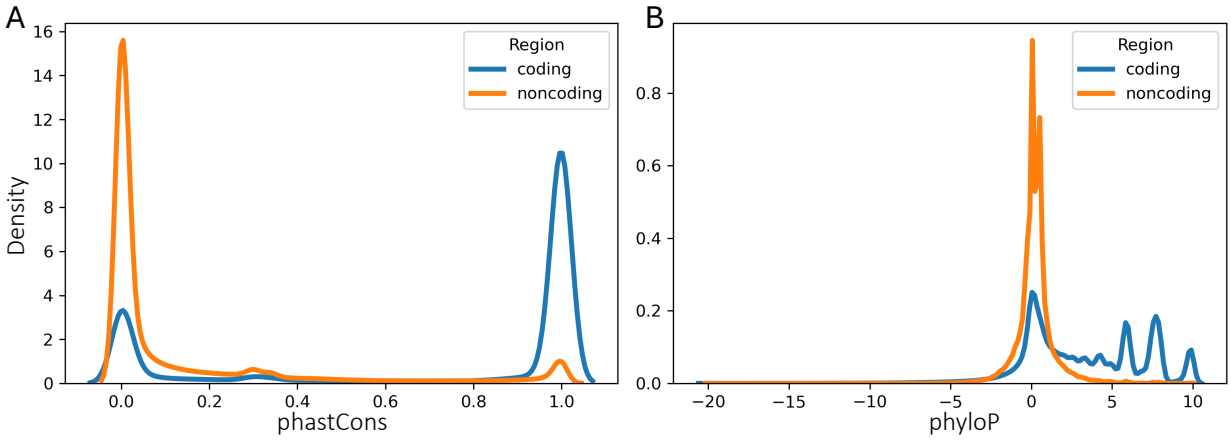
adjustments. An incremental learning strategy was adopted to leverage human data for cattle variant prediction (Wu et al., 2019). The human model, trained on human annotations available for both human and cattle datasets, was employed as the initial model for a subsequent training step utilizing cattle annotations. For the comparison of models with and without Enformer features, we trained these models using the CatBoost algorithm with default hyperparameters due to limitations in available computational resources.

To evaluate the models' performance beyond just accuracy, various metrics were calculated using scikit-learn packages, including precision, recall, F1-score, and the Area Under the Curve (AUROC), to assess their proficiency in true positive predictions and discriminative capacity between regulatory variants and other variants. Furthermore, the Shapley Additive exPlanations (SHAP) approach was employed to elucidate the relationship between feature values and model outputs, thereby enhancing model interpretation (Lundberg & Lee, 2017).

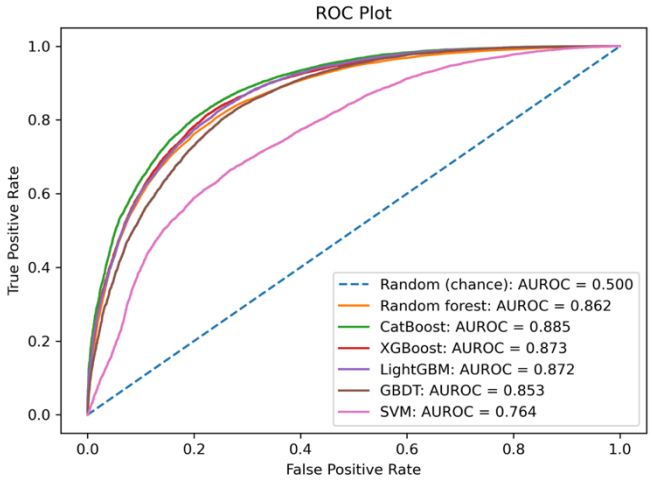
Acknowledgments

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). This project was also supported by funding by the Bill and Melinda Gates Foundation (BMGF) and with UK aid from the UK Government's Department for International Development (Grant Agreement OPP1127286) under the auspices of the Centre for Tropical Livestock Genetics and Health (CTLGH), established jointly by the University of Edinburgh, SRUC (Scotland's Rural College), and the International Livestock Research Institute (ILRI). The findings and conclusions are those of the authors and do not necessarily reflect the positions or policies of the BMGF or the UK Government.

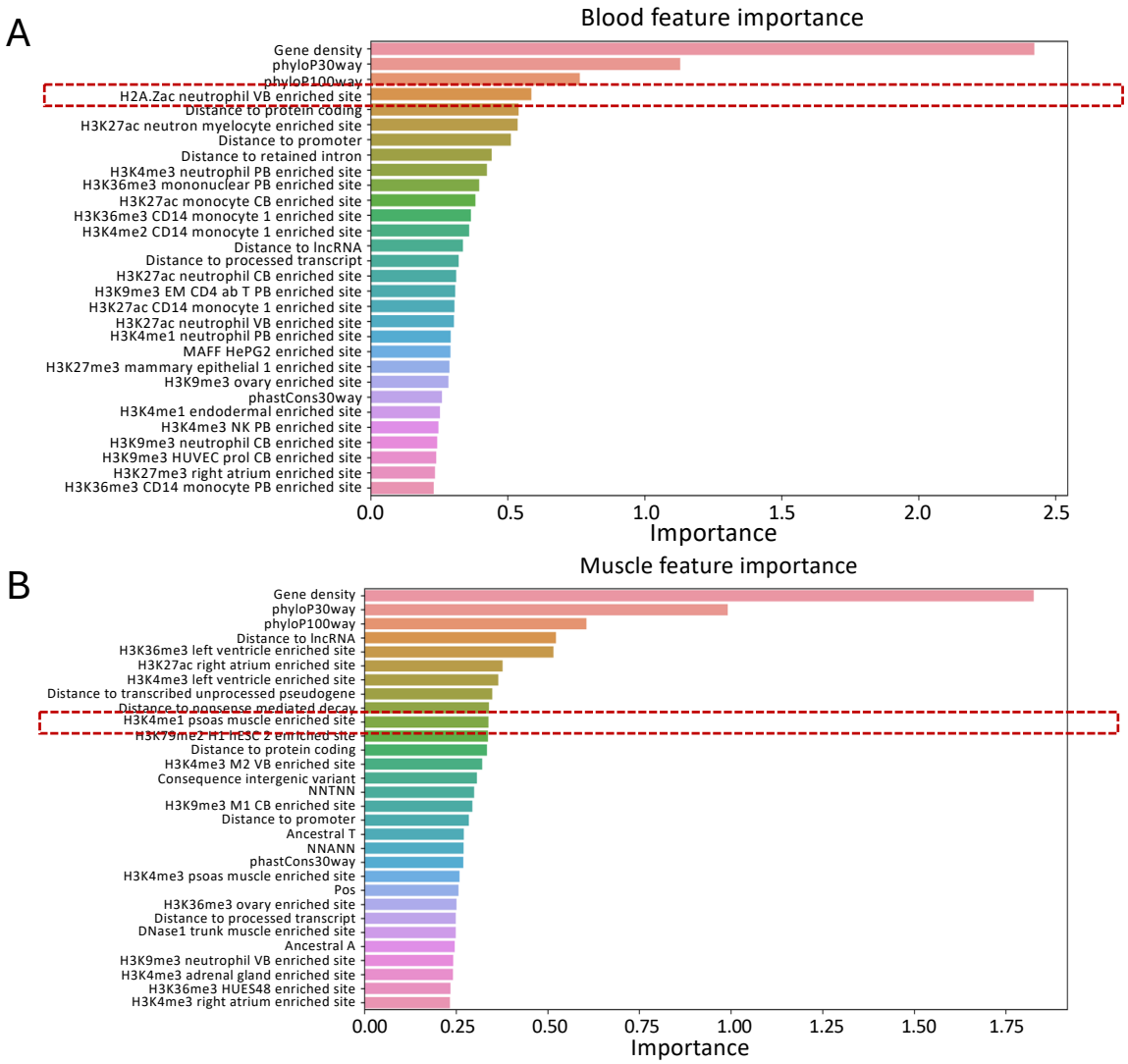
Supplementary materials



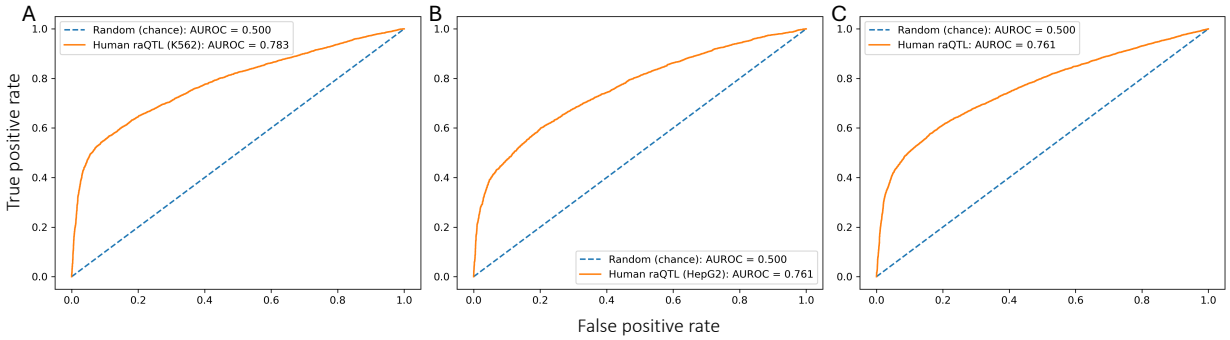
Supplementary Figure 1 Conservation score distribution differences in coding and noncoding regions in the cattle genome. (A) phastCons241way conservation score (B) phyloP241way conservation score.



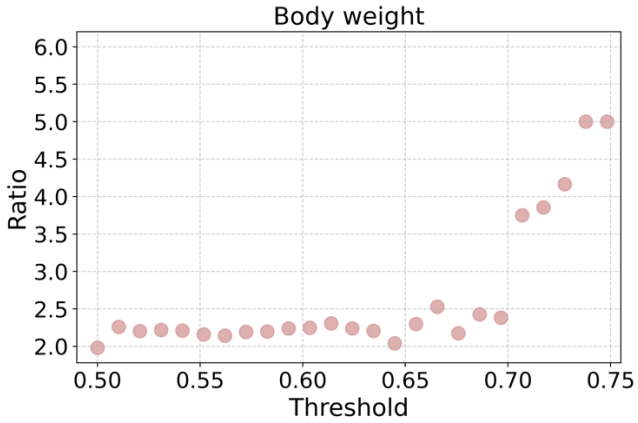
Supplementary Figure 2 ROC curves and AUROC scores for seven machine learning models trained and tested using the high-confidence human data.



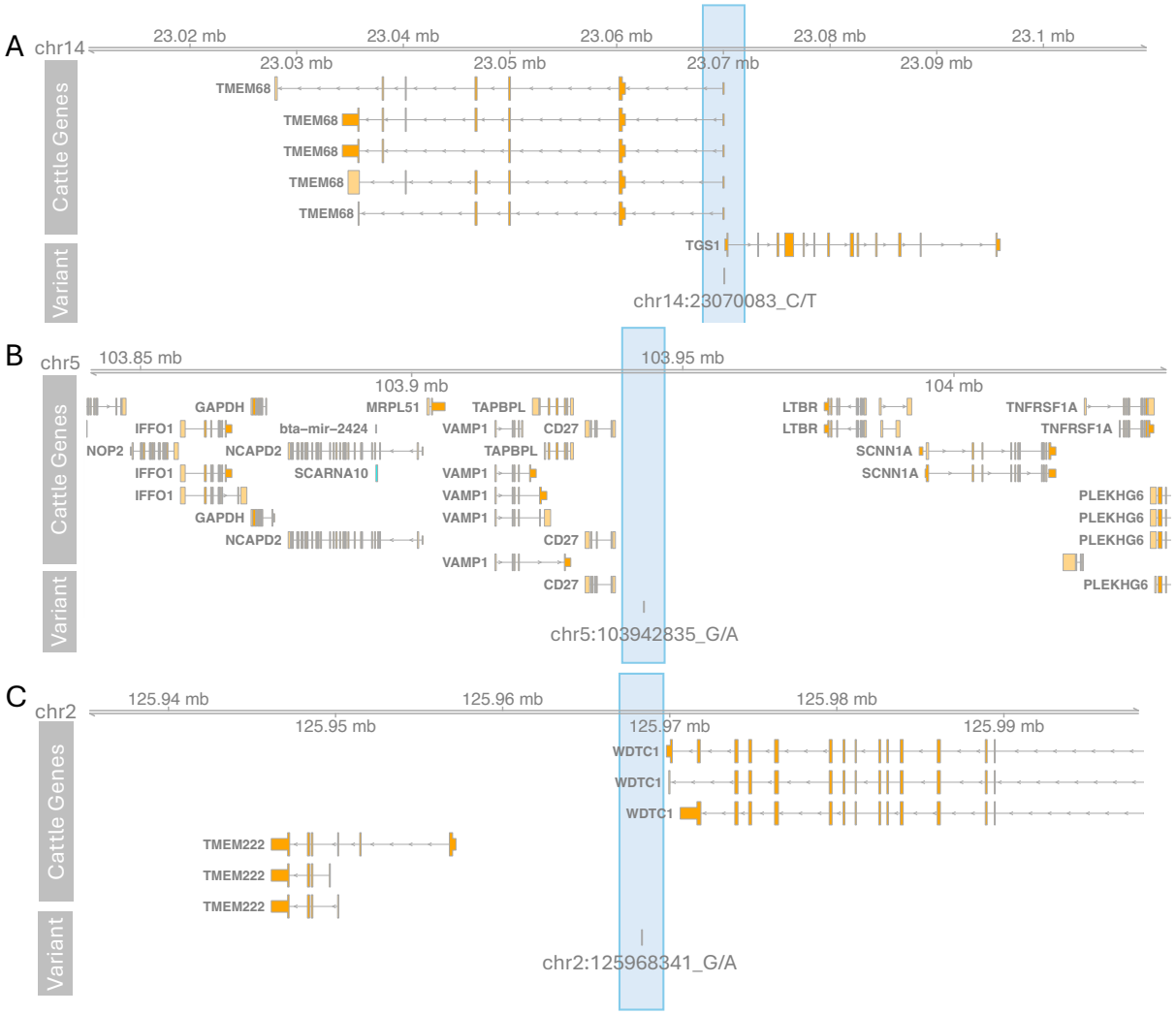
Supplementary Figure 3 Top 30 features in blood and muscle specific models. The features are ranked in descending order according to their relative feature importance in the model. The features within the red dashed rectangle are the top chromatin features that exclusively emerged among the top 10 in their respective tissue-specific models.



Supplementary Figure 4 ROC plots and AUROC scores for human raQTL models. (A) in K562 cell line. (B) in HepG2 cell line. (C) in both cell lines.



Supplementary Figure 5 Different prediction probability thresholds and their corresponding ratios of the proportion of cattle GWAS fine-mapped variants associated with body weight predicted as regulatory to the proportion of background variants predicted as regulatory using the cross-species model.



Supplementary Figure 6 Three cattle trait-associated variants consistently predicted as regulatory variants by all models. (A) stature-associated variant chr14:23070083_C/T, (B) body weight-associated variant chr5:103942835_G/A, and (C) body condition-associated variant chr2:125968341_G/A.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. https://doi.org/10.1038/nature15393

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

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

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

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), 1747–1751. https://doi.org/10.1038/ng.3979

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience, 4(1), s13742-015-0047–0048. https://doi.org/10.1186/s13742-015-0047-8

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, T., & Guestrin, C. (2016). 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

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

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

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

Ducos, E., Vergès, V., Dugé de Bernonville, T., Blanc, N., Giglioli-Guivarc'h, N., & Dutilleul, C. (2017). Remarkable Evolutionary Conservation of Antiobesity ADIPOSE/WDTC1 Homologs in Animals and Plants. Genetics, 207(1), 153–162. https://doi.org/10.1534/genetics.116.198382

Englebert, K., Taquin, A., Azouz, A., Acolty, V., Vande Velde, S., Vanhollebeke, M., Innes, H., Boon, L., Keler, T., Leo, O., Goriely, S., Moser, M., & Oldenhove, G. (2024). The CD27/CD70 pathway negatively regulates visceral adipose tissue-resident Th2 cells and controls metabolic homeostasis. Cell Reports, 43(3), 113824. https://doi.org/10.1016/j.celrep.2024.113824

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

Genereux, D. P., Serres, A., Armstrong, J., Johnson, J., Marinescu, V. D., Murén, E., Juan, D., Bejerano, G., Casewell, N. R., Chemnick, L. G., Damas, J., Di Palma, F., Diekhans, M., Fiddes, I. T., Garber, M., Gladyshev, V. N., Goodman, L., Haerty, W., Houck, M. L., ... Zoonomia Consortium. (2020). A comparative genomics multitool for scientific discovery and conservation. Nature, 587(7833), 240–245. https://doi.org/10.1038/s41586-020-2876-6

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

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

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

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

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

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

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

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

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

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.

Kern, C., Wang, Y., Xu, X., Pan, Z., Halstead, M., Chanthavixay, G., Saelao, P., Waters, S., Xiang, R., Chamberlain, A., Korf, I., Delany, M. E., Cheng, H. H., Medrano, J. F., Van Eenennaam, A. L., Tuggle, C. K., Ernst, C., Flicek, P., Quon, G., ... Zhou, H. (2021). Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nature Communications, 12(1), 1821. https://doi.org/10.1038/s41467-021-22100-8

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

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

Lee, Y.-L., Takeda, H., Moreira, G. C. M., Karim, L., Mullaart, E., Coppieters, W., Consortium, T. G., Appeltant, R., Veerkamp, R. F., Groenen, M. A. M., Georges, M., Bosse, M., Druet, T., Bouwman, A. C., & Charlier, C. (2021). A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy cattle. PLOS Genetics, 17(7), e1009331. https://doi.org/10.1371/journal.pgen.1009331

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

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

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., ... Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. Nature Genetics, 45(6), 580–585. https://doi.org/10.1038/ng.2653

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

Martin, F. J., Amode, M. R., Aneja, A., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Bignell, A., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Charkhchi, M., Cockburn, A., Da Rin Fiorretto, L., ... Flicek, P. (2023). Ensembl 2023. Nucleic Acids Research, 51(D1), D933–D941. https://doi.org/10.1093/nar/gkac958

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

Mostafavi, H., Spence, J. P., Naqvi, S., & Pritchard, J. K. (2023). Systematic differences in discovery of genetic effects on gene expression and complex traits. Nature Genetics, 55(11), 1866–1875. https://doi.org/10.1038/s41588-023-01529-1

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

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

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.

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

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

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

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). 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

Reynolds, E. G. M., Neeley, C., Lopdell, T. J., Keehan, M., Dittmer, K., Harland, C. S., Couldrey, C., Johnson, T. J. J., Tiplady, K., Worth, G., Walker, M., Davis, S. R., Sherlock, R. G., Carnie, K., Harris, B. L., Charlier, C., Georges, M., Spelman, R. J., Garrick, D. J., & Littlejohn, M. D. (2021). Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. Nature Genetics, 53(7), 949–954. https://doi.org/10.1038/s41588-021-00872-5

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

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., ... Sherry, S. T. (2022). Database

resources of the national center for biotechnology information. Nucleic Acids Research, 50(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

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

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

Smedley, D., Schubach, M., Jacobsen, J. O. B., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N. L., McMurry, J. A., Haendel, M. A., Mungall, C. J., Lewis, S. E., Groza, T., Valentini, G., & Robinson, P. N. (2016). A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. The American Journal of Human Genetics, 99(3), 595–606. https://doi.org/10.1016/j.ajhg.2016.07.005

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

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

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

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

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

Wong, E. S., Zheng, D., Tan, S. Z., Bower, N. I., Garside, V., Vanwalleghem, G., Gaiti, F., Scott, E., Hogan, B. M., Kikuchi, K., McGlinn, E., Francois, M., & Degnan, B. M. (2020). Deep conservation of the enhancer regulatory code in animals. Science, 370(6517), eaax8137. https://doi.org/10.1126/science.aax8137

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., Berg, I. van den, MacLeod, I. M., Hayes, B. J., Prowse-Wilkins, C. P., Wang, M., Bolormaa, S., Liu, Z., Rochfort, S. J., Reich, C. M., Mason, B. A., Vander Jagt, C. J., Daetwyler, H. D., Lund, M. S., Chamberlain, A. J., & Goddard, M. E. (2019). Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. Proceedings of the National Academy of Sciences, 116(39), 19398–19408. https://doi.org/10.1073/pnas.1904159116

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., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A. F., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., Frayling, T. M., McCarthy, M. I., Hirschhorn, J. N., Goddard, M. E., & Visscher, P. M. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics, 44(4), 369-S3. https://doi.org/10.1038/ng.2213

Zhao, F., McParland, S., Kearney, F., Du, L., & Berry, D. P. (2015). Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genetics Selection Evolution, 47(1), 49. https://doi.org/10.1186/s12711-015-0127-3

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, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S., & Green, M. R. (2010). ChIPpeakAnno: A Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics, 11(1), 237. https://doi.org/10.1186/1471-2105-11-237