The potential of AI models to identify regulatory variants underlying cattle traits

Zhao R, Owen R, Marr M, Jayaraman S, 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 0.86 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 18-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 play a critical role in shaping mammalian phenotypes and are estimated to account for approximately 60% of the genetic differences underlying human disease (Nica & Dermitzakis, 2013) and around 70% of the heritability for many cattle phenotypes (Xiang et al., 2023). Genome-wide association studies (GWAS) have successfully linked numerous genomic regions to various diseases and traits in both humans and livestock; however, pinpointing the specific causal variants within these regions remains challenging. One major obstacle is linkage disequilibrium (LD) among nearby variants, that is particularly high in livestock breeds, and which makes it difficult to isolate the genetic changes driving the trait. Although intersecting GWAS findings with known expression quantitative trait loci (eQTL) helps suggest regulatory variation as a plausible cause of a GWAS peak, the precise causal variant still often remains elusive.

To overcome this challenge there has been growing interest in developing artificial intelligence (AI) models for predicting functional regulatory variants, though primarily in the human genome. One early example of such an AI model was GWAVA (Ritchie et al., 2014), which employed a random forest model trained on regulatory variants sourced from the human gene mutation database (HGMD) (Stenson et al., 2020). GWAVA leveraged a broad set of genomic features, including chromatin data and sequence conservation, to classify new variants, with an area under the receiver operating characteristic curve (AUROC), a combined measure of the model's performance across classification thresholds, ranging from 0.75 to 0.84 depending on the chosen background set of non-functional variants. DeepSEA (Zhou & Troyanskaya, 2015) used a convolutional neural network (CNN) to predict chromatin profiles from local sequence and demonstrated that these predictions could be used to classify regulatory variants. Using the predicted effects of sequence changes on chromatin profiles, its models achieved an AUROC of up to 0.7 and outperformed GWAVA when both were trained on the same HGMD data. More recently, DeepSEA's foundational work was extended to include 21,907 chromatin profiles covering over 1,300 cell lines and tissues, culminating in the Sei model (K. M. Chen et al., 2022). Mutations in HGMD had Sei functional effect scores

up to 6.5 times higher than background variants. Another key development is the Enformer model (Avsec et al., 2021), which, similar to DeepSEA, leverages CNNs to predict large-scale omics data from sequence. Training random forest classifiers on Enformer output to distinguish fine-mapped GTEx (GTEx Consortium, 2020) regulatory variants from background variants yielded an AUROC of 0.75.

Despite these advances, most of these efforts to predict functional regulatory variants have focused on humans due to the abundant availability of omics data. Yet extending similar methods to livestock species is of considerable importance. Selective breeding programs in livestock could potentially incorporate functional variant predictions to accelerate genetic gains and yield tangible societal benefits more rapidly than would be feasible for many other species. For example, previous cattle studies found that prioritising variants based on functional-and-evolutionary trait heritability (FAETH) scores (Xiang et al., 2019), encompassing a range of regulatory and evolutionary features, had the potential to enhance the rate of genetic improvement for economically important phenotypes (Xiang et al., 2021). Simulation studies further support the idea that enriching for functional variants can help improve genomic estimated breeding values (Santana et al., 2023). Consequently, identifying functional variants in livestock has the possibility to spur both economic and global public health advances.

A key hurdle to developing advanced machine learning models for livestock species is the scarcity of high-quality training data. In humans, the genotype-tissue expression (GTEx) project has become a gold-standard resource for mapping gene expression and refining genomic regions to fine-mapped regulatory variant using tools such as CaVEMaN (Brown et al., 2017). Inspired by GTEx, the farm animal genotype-tissue expression (FarmGTEx) project aims to establish similarly comprehensive resources in key farm animals, including pigs and cattle (Liu et al., 2022; Teng et al., 2024). For cattle specifically, the FarmGTEx project analysed 7,180 publicly available RNA-Seq samples derived from 46 breeds (or breed combinations) across 114 distinct tissues. eQTL analyses were subsequently carried out for 23 of these tissues. While these collective datasets have significantly advanced eQTL mapping in livestock, challenges remain. For example, the

FarmGTEx cattle resource had to integrate heterogenous public datasets generated by different labs at different times, and variant calling was performed solely on RNA-Seq data - ultimately limiting the number of variants detected.

Another major challenge in applying machine learning approaches to livestock species is the relative paucity of specialised functional annotation tools and datasets. Variant annotation is central to genomic analysis, providing functional context to DNA variants. Widely used resources include ANNOVAR (Wang et al., 2010), based on the UCSC Genome Browser database, and the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016). Although these tools can deliver a wealth of annotation metadata, they often lack species-specific or downstream-analysis–oriented features, especially those relevant to machine learning approaches for regulatory variant prediction. They can also be human-centric, further underscoring the need for annotation frameworks that are readily adaptable to livestock. Likewise, conservation-based metrics such as PhyloP and PhastCons (Pollard et al., 2010), that have been shown to be invaluable for uncovering functional elements in the human genome (Sullivan et al., 2023), are not currently available for cattle and many other farm animals.

In this study, we sought to assess the feasibility of developing machine learning models that predict regulatory variants in livestock, with a particular focus on cattle. Our first step was to create a species-agnostic variant annotation pipeline tailored for developing machine learning models. This included deriving key PhyloP and PhastCons conservation metrics for cattle for the first time. We then benchmarked prospective modelling approaches by first applying them to human data and comparing that performance to analogous models in cattle. This comparison allowed us to identify potential performance gaps and clarify which new datasets and methods may be most critical for future improvements in cattle regulatory variant prediction. We investigated the effectiveness of different training approaches 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 underlying economically and clinically relevant traits.

Results

Novel variant annotation pipeline and conservation metrics

To overcome the limitations of existing variant annotation tools for non-human species, we developed *nf-VarAnno*; a species-agnostic and reusable variant annotation pipeline built with Nextflow (Di Tommaso et al., 2017) (Figure 1; https://github.com/evotools/nf-VarAnno). As summarised in Table 1, the pipeline offers five categories of annotations: (1) sequence conservation metrics, (2) variant position properties, (3) VEP annotations, (4) sequence context and (5) predicted functional genomic scores derived from the Enformer deep learning sequence-based model (Avsec et al., 2021).

Among these annotation types, sequence conservation metrics are particularly powerful annotations for identifying functional variation in the genome (Sullivan et al., 2023; Xiang et al., 2019). However, aside from GERP (Davydov et al., 2010), most major conservation scores are not available for livestock. To address this gap, we calculated 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 Fig. 1A and B, these newly derived metrics are enriched in coding regions, validating their ability to identify functional genomic regions. We have made the full set of cattle conservation metrics publicly available at Zenodo (https://zenodo.org/doi/10.5281/zenodo.13332540).

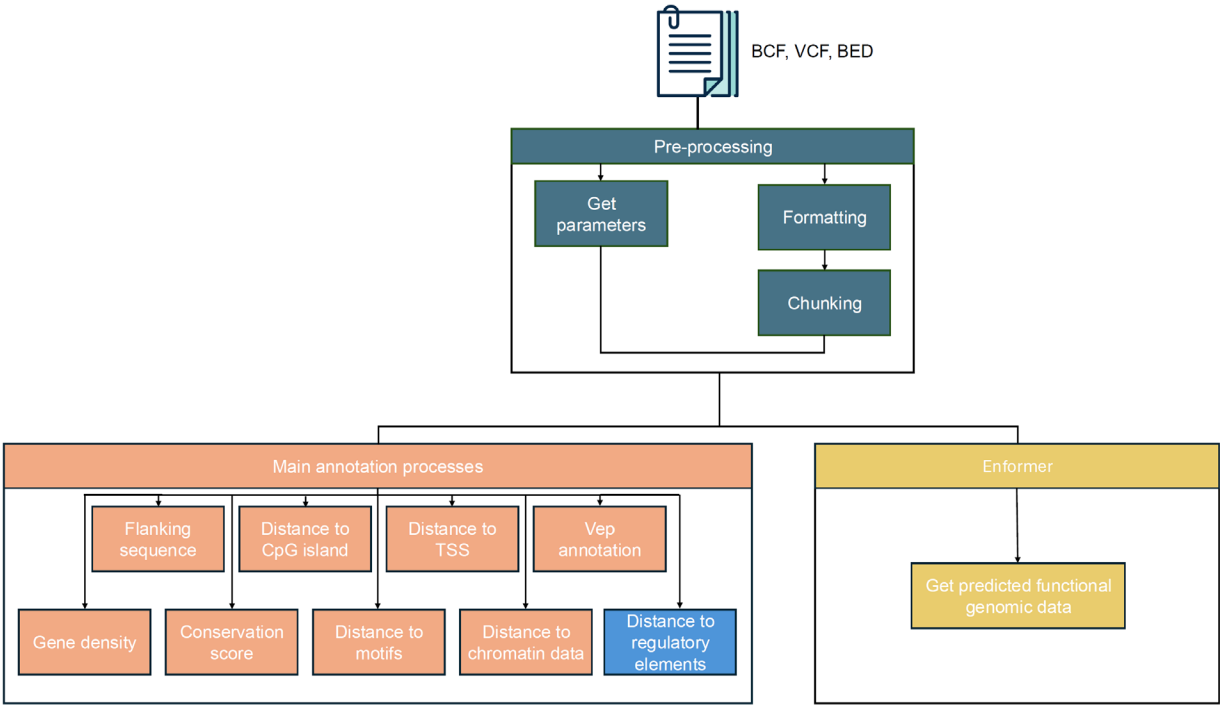


Figure 1. The workflow for the annotation pipeline. The workflow consists of three sections, pre-processing, the main annotation sub-workflow, and Enformer sub-workflow. The blue rectangles indicate annotations specific to humans.

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	Human: 1554, Cattle: 95	
	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	
	Distance to motifs	-	HOMER, Bedtools	460	
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

153 **The conservation of regulatory variant annotations across species**

154 We next applied our species-agnostic variation annotation pipeline to both human and
155 cattle datasets. In humans, we started with 590,778 regulatory variants identified through
156 CaVEMaN fine-mapping of GTEx data, among which 45,987 exhibited a CaVEMaN
157 causal probability greater than 0.5 and were classified as high-confidence regulatory
158 variants. For cattle, we obtained 79,215 cis-regulatory variants from the cattle FarmGTEx
159 project. To facilitate a direct comparison of annotations between regulatory and
160 background variants, we randomly sampled a background set of variants in each species
161 within 1Mb of transcription start sites (TSS) – consistent with the GTEx cis-eQTL detection
162 window – and then matched their minor allele frequency (MAF) distributions to those of
163 the foreground regulatory variants (See Methods).
164 Figure 2 demonstrates that many annotation features exhibit similar associations with
165 regulatory variants across species. For example, regulatory variants in both humans and
166 cattle are more likely to involve a G to A (C to T) allele change (Figure 2A, B), and the
167 flanking sequences have a tendency to harbor C:G base pairs (Figure 2I,J). Furthermore,
168 regulatory variants are more abundant in gene-dense regions and exhibit similar
169 enrichments in proximity to specific genomic elements, such as CpG islands and
170 transcription start sites (TSS) (Figure 2C to H). Notably, they also tend to occupy more
171 evolutionarily conserved genomic sites, on average, than the background variants
172 (Supplementary Fig. 2). Taken together, these observations highlight key commonalities
173 in the genomic and evolutionary features of regulatory variants across different species.
174
175
176
177
178
179
180

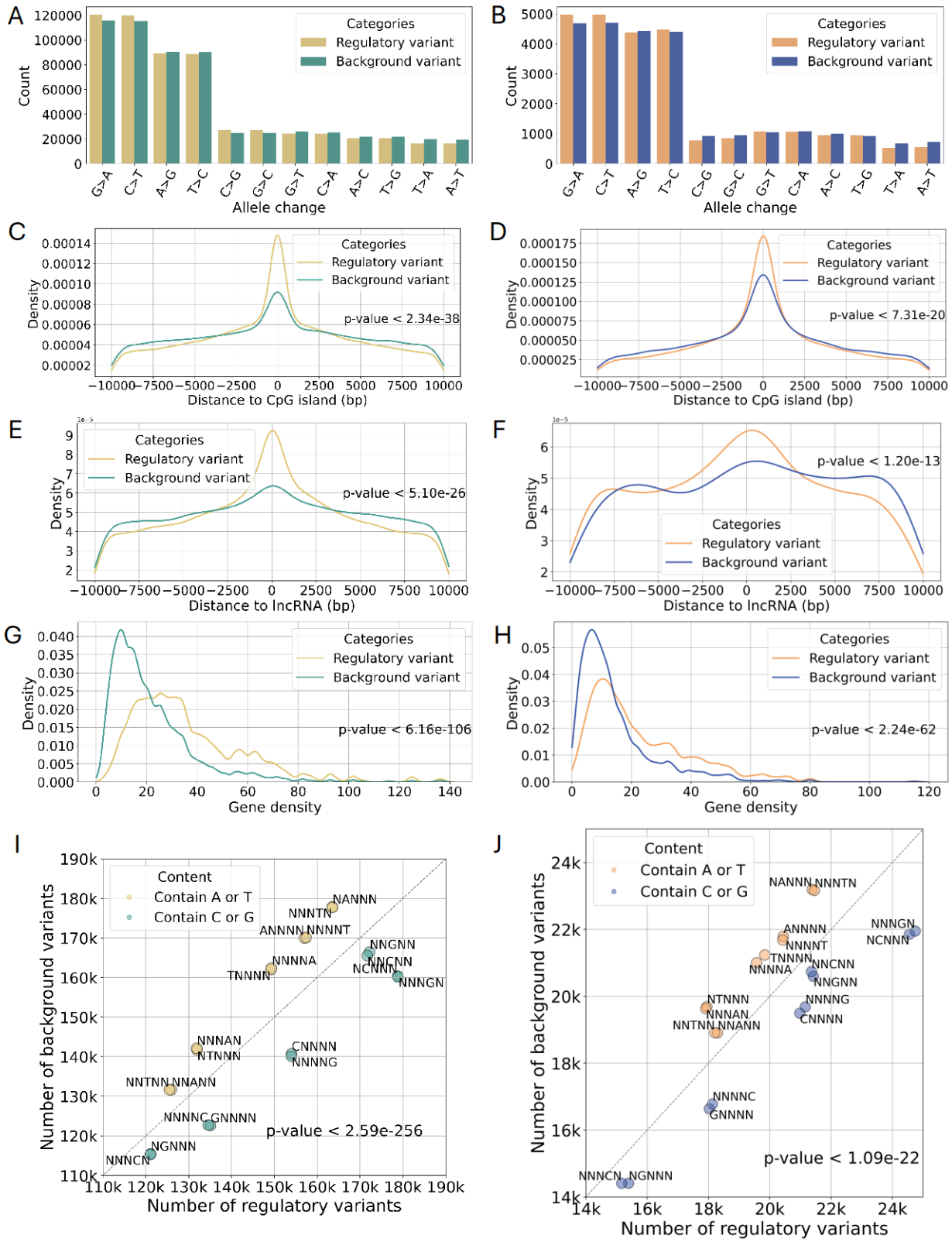


Figure 2 Common characteristics of human and cattle regulatory variants. Number of regulatory and background variants by their allele changes in (A) human and (B) cattle. Kernel density estimate (KDE)

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. A matching set of background variants was selected from the remaining variants tested in GTEx (See methods). The different sets of variants used in this study can be found in Supplementary Fig. 3. These variants were annotated with general features as indicated in Table 1, excluding the Enformer features that are discussed in a later section. The models were evaluated using a chromosome-level holdout strategy. Specifically, certain chromosomes were set aside as test sets to assess model performance while the remaining chromosomes were used for hyper-parameter tuning (See Methods). Here, we present the model performance tested on chromosomes 1 and 22 (chr1 and chr22). Other test results can be found in Supplementary Fig. 4. The key feature distributions between foreground and background data in the training and test sets are shown in Supplementary Fig. 5. As illustrated in Supplementary Fig. 5, the training and test sets exhibit similar feature distributions for foreground and background data. Different machine learning strategies, including Random Forest, CatBoost, XGBoost, and SVM, were explored, achieving test AUROC scores of 0.82, 0.86, 0.83 and 0.64, respectively, at distinguishing fine-mapped regulatory variants from matching background variants. Considering both its low training time and

higher accuracy, we consequently selected the CatBoost model as the algorithm for all downstream analyses.

Leveraging the TraitGym (Benegas, Eraslan, et al., 2025) curated set of putative functional regulatory variants linked to human complex traits, we further benchmarked how well this modelling approach and set of features distinguish human regulatory changes from background variants. The resulting performance (average test AUROC of 0.67 with a standard deviation of 0.04) was broadly comparable to that of the latest version of CADD (~0.69) (Schubach et al., 2024) the best performing model on this dataset in a recent benchmark (Benegas, Eraslan, et al., 2025), despite CADD's reliance on a far more extensive set of features - many of which that would be hard to obtain for training livestock models. Performance was also comparable to the recently released GPN-MSA DNA language model (AUROC of ~0.66) (Benegas, Albors, et al., 2025). Notably, these results were obtained even though one of the potentially most informative biological attributes for classifying regulatory variants, distance to promoters, is effectively controlled for in the TraitGym dataset with the foreground and background variant sets matched on this feature.

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 matching background variants were down-sampled 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). Among all models, the high-confidence model achieved the highest test AUROC of 0.86 (Figure 3A), and the highest mean cross-validation AUROC score of 0.87 (Supplementary Fig. 4A). To assess the robustness of the high-confidence model, we repeated training and testing 10 times using different randomly selected matching background sets. The models yielded

248 test AUROCs ranging from 0.85 to 0.87, with a mean test AUROC of 0.86 and a standard
249 deviation of 0.006. Although direct comparisons to the performance of other models
250 trained on different datasets should be done with caution, these results compare
251 favourably to other published human models (Avsec et al., 2021; Ritchie et al., 2014; Zhou
252 & Troyanskaya, 2015) and highlights the utility of our workflow for training models that
253 substantially enrich for regulatory variants.
254
255 To understand which features impacted prediction outcomes, and therefore may be most
256 relevant to training cattle models, a SHAP summary plot was generated for model
257 interpretability (S. M. Lundberg & Lee, 2017). A SHAP summary plot visualizes the impact
258 of different features on a machine learning model's prediction, offering insights into
259 feature importance and the direction of their influence (S. M. Lundberg & Lee, 2017). The
260 SHAP values in the summary plot quantify each feature's contribution to the model's
261 predictions, indicating both the magnitude and direction of their impact on individual data
262 points. Figure 3B illustrates the top 30 most important features for the best-performing
263 cross-tissue model, along with their corresponding impacts on the model's predictions.
264 Note the annotation categories described in Table 1 can be split across multiple features
265 after feature encoding, such as the VEP variant effect predictions. The most influential
266 feature is the distance to the promoter. As shown in the plot, a moderate distance-to-
267 promoter value (around 0, as distance includes both positive and negative values),
268 indicating proximity to the promoter region, is linked to a higher probability of a variant
269 being predicted as a regulatory variant. Other features that prominently contribute to the
270 model outcomes include gene density, variant consequence, distance to TSS, and
271 conservation scores. In the tissue-specific models, in addition to features that exhibit
272 importance across different models, such as gene density and distance to TSS, certain
273 chromatin features emerged solely among the top features in specific tissue models. For
274 example, the distance from the variant to open chromatin regions marked by H3K4me3
275 in the psoas muscle featured among the top 10 most important attributes in the muscle-
276 specific model (Supplementary Fig. 6A). Similarly, the distance from the variant to open
277 chromatin regions marked by H3K4me3 in the neutrophil cell line ranked among the top
278 5 most important features exclusively in the blood-specific model (Supplementary Fig.

6B). This suggests that tissue-specific chromatin data can, to some extent, improve tissue-specific model predictions.

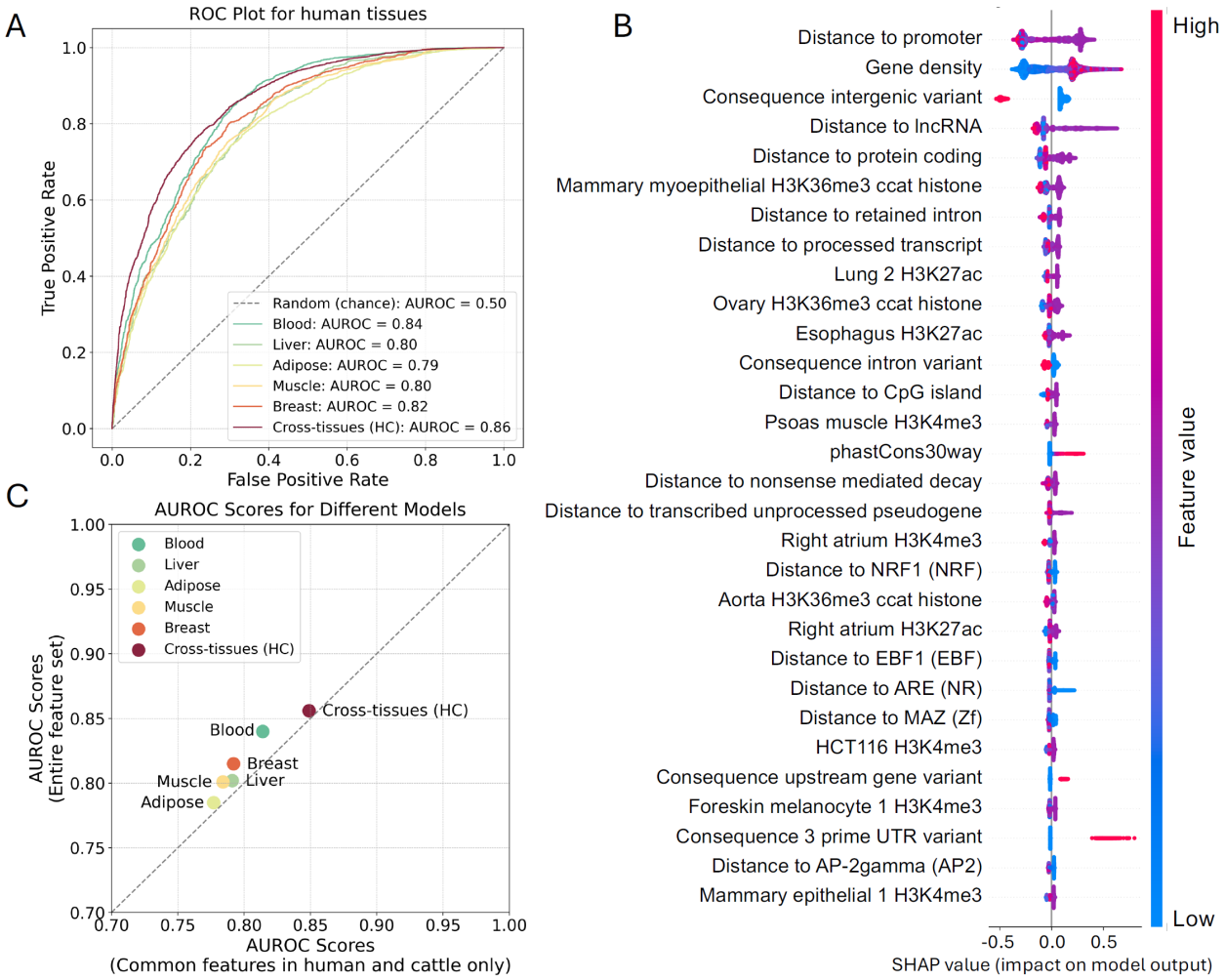


Figure 3 Human model performances. (A) Receiver Operating Characteristic (ROC) curves for human tissue-specific and cross-tissues high-confidence (HC) model. The numbers in the plot represent the area under the receiver operating characteristic (AUROC) scores for different models (ranging from 0 to 1), 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 shows 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 0.70 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, many of the most 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 ~0.03), their impact on cross-tissues models is minimal (only ~0.007 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. The cattle models were also evaluated using a chromosome-level holdout strategy (See Methods). Here, we present the model performance tested on chromosome 1 and 28 (chr1 and chr28). The key feature distributions of foreground and background variants in the training and test sets are shown in Supplementary Fig. 5. Results tested on other chromosomes can be found in Supplementary Fig. 7. The cross-tissue cattle model achieved a test AUROC score of 0.73 (Figure 4A), and a mean cross-validation AUROC score of 0.71 (Supplementary Fig. 7A). We repeated training and testing 10 times using different randomly selected matching background sets to assess the robustness of cattle cross-tissues model. The models yielded test AUROCs ranging from 0.71 to 0.74, with a mean test AUROC of 0.73 and a standard deviation of 0.01. Comparison of the feature importances for these human and cattle cross-tissue models trained on the common sets of over 470 features showed that the most important features determining the prediction results were largely consistent across the two models, with the top seven 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, the muscle model achieved the highest test AUROC score of 0.75 (Figure 4A) (mean cross-validation AUROC score of 0.76). In general, cattle models achieved poorer performance compared to human models when trained on the equivalent set of features.

To explore whether incorporating cattle chromatin data could improve model performance, we included 95 cattle chromatin datasets. The features were used both in a cross-tissue manner and in a tissue-specific manner for available tissues (liver, adipose, and muscle). However, the model's performance remained the same (Supplementary Fig. 8), suggesting the addition of the cattle chromatin data had little improvement on the accuracy of predicting regulatory variants.

Given the difference in the number of human and cattle variants available for training, we down-sampled the human variant sets to match the number of variants in cattle models for a fair comparison. Despite this adjustment, the human models still performed better (Supplementary Fig. 9). These observations suggest that the performance gap between species is potentially largely driven by differences in the quality of the regulatory variant sets used, rather than limitations in the available feature sets.

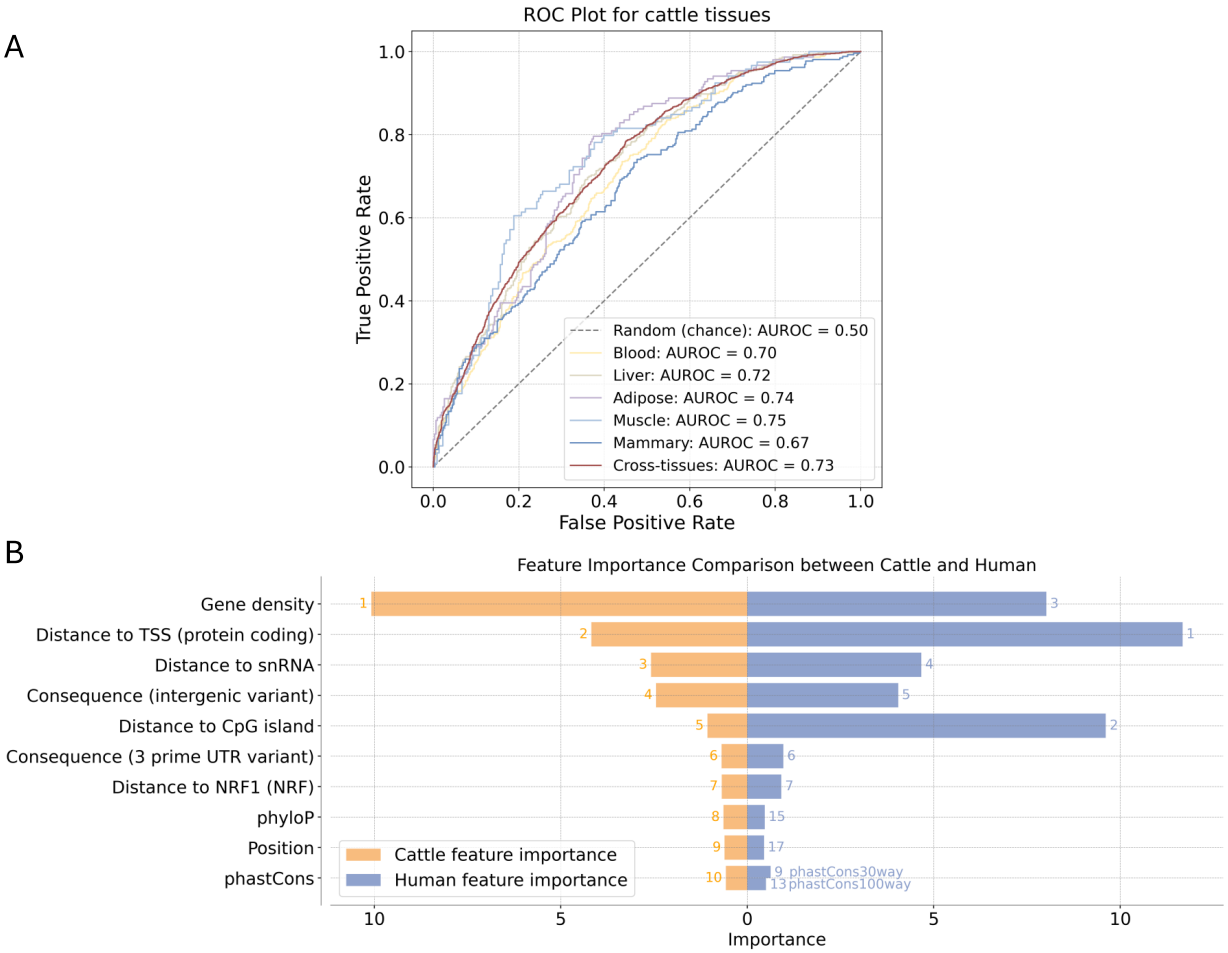


Figure 4 The comparison of cattle and human models trained on the common set of features. (A) ROC curves for cattle tissue-specific and cross-tissues models. The numbers in the plot legend represent the AUROC scores of different models. (B) Bar plot showing the top features in the cattle (orange) and 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 phastCons conservation scores were available (phastCons30way and phastCons100way), ranking 9th and 13th.

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

359 based on the human model for cattle prediction, and lifting cattle variants to the human
360 genome to train the model using human annotations. Note that in the latter only about 52%
361 of the cattle GTEx variants could be lifted over to the human genome. These approaches
362 achieved test AUROC scores of 0.67, 0.68, and 0.68, respectively, which are lower than
363 those of the cattle models trained on all cattle annotations except for cattle chromatin data
364 (AUROC = 0.73). Therefore, this outcome highlights that species-specific models
365 outperform, even when fewer annotations are available. Human-based models, while
366 capturing some general principles, may miss species-specific intricacies crucial for
367 accurate predictions.
368

369 **Enformer provides little benefit in predicting regulatory variants in either species**

370 Enformer is a sequence-based model that can predict a range of regulatory features from
371 sequence alone, and the authors illustrated its potential at predicting regulatory variants
372 by comparing regulatory feature predictions when variant alleles are changed in the input
373 sequence (Avsec et al., 2021). Such models therefore provide a potential avenue to
374 enhance feature sets from just genome sequences, and if it could, at least in part, be
375 applied across species, it would have the potential to fill in gaps in omics data for less
376 well characterised mammals. We consequently investigated the usefulness of features
377 derived from the sequence-based Enformer model to inform both human and cattle
378 predictions. The human high-confidence cross-tissues dataset was annotated with 5,313
379 Enformer features, i.e. the difference in predicted omics data between genomic
380 sequences with reference and alternative alleles. We then trained CatBoost models using
381 three feature sets: Enformer features only, general features only, and a combination of
382 both. As shown in Figure 5A, the model trained with general features performed slightly
383 better than the one using only Enformer features (AUROC score increase of 0.02). The
384 combined model achieved a slight improvement in performance, with an AUROC score
385 increase of 0.01 compared to the general features model. This suggests that while
386 incorporating Enformer features may provide some limited additional regulatory
387 information, its overall contribution to predictive power remains modest. We further
388 applied the Enformer sub-workflow to annotate cattle tissue-specific variants. This aimed
389 to explore whether these inferred chromatin features could improve cattle model

predictions, given the comparative lack of omics data for this species. We initially applied the full set of Enformer features but the performance remained unchanged compared to not using Enformer features (Supplementary Fig. 10). Subsequently we incorporated tissue-relevant Enformer features into the cattle tissue-specific models, which resulted in a modest improvement in performance, with a maximum AUROC score increase of 0.005 (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).

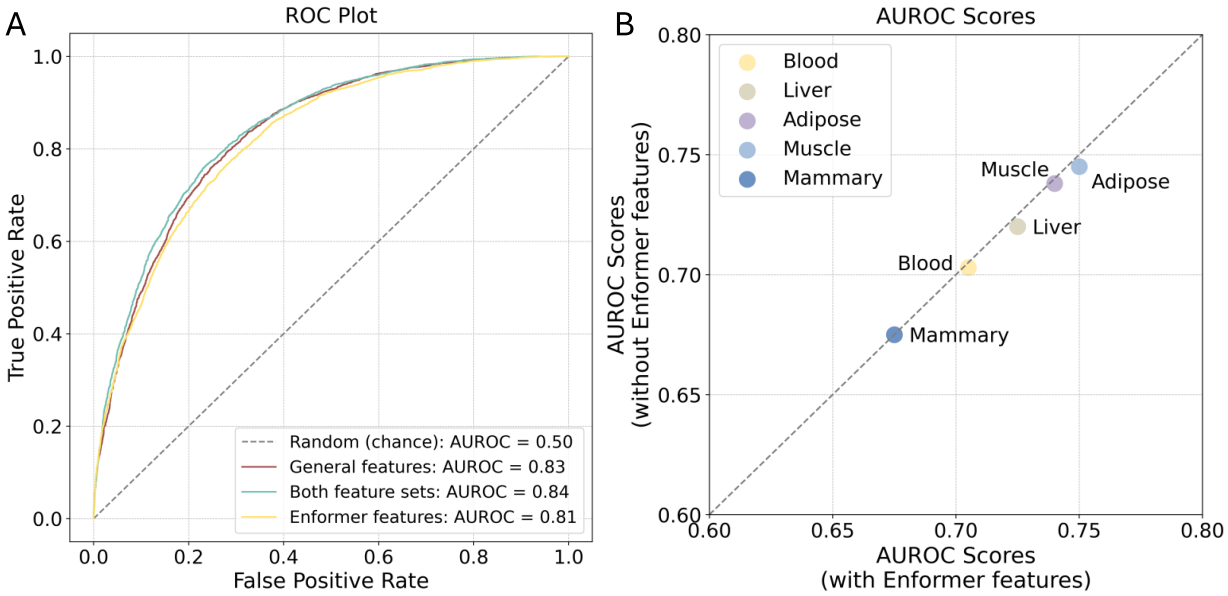


Figure 5 The comparison of human and cattle models trained with or without Enformer inferred features. (A) ROC curves for human high-confidence cross-tissues model trained using Enformer features, general features, and the combination of two feature sets. The numbers in the plot represent the AUROC scores of different models. (B) The comparison of the AUROC scores of cattle tissue-specific models trained with or without tissue-specific 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 evaluate the effectiveness of our models at predicting regulatory variants across diverse genomic regions, we compared their performance on variants stratified into five distance ranges relative to TSSs in different tissues. For each distance bin, the major

class (i.e., the class with more variants) was down-sampled to match the number of the minor class, creating a balanced set. Bins were defined to ensure comparable variant counts after down-sampling within each tissue. In cattle, tissues with fewer than 10,000 variants were excluded. As illustrated in Figure 6, model performance displayed a slight downward trend with increasing distance from the TSS. Despite this, the overall stability of performance, even for distal variants, suggests that our model generalizes effectively across the genome, including regions far from canonical regulatory hubs.

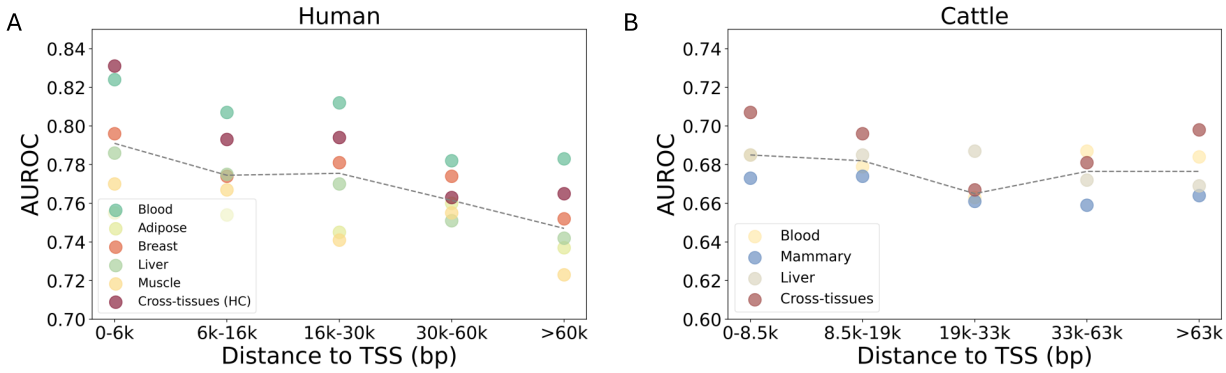


Figure 6 Human and cattle model performance using variants within different distance ranges to the TSS. (A) Human tissue-specific and cross-tissues high-confidence (HC) models performance based on variants in different TSS bins. The grey dashed line is plotted based on the median test AUROC score among the six datasets in each range. (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 test AUROC 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 eQTLs. 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.

437 Massively parallel reporter assays (MPRA) are an alternative approach for mapping
438 regulatory variation in the genome that overcome some of these limitations. For example,
439 approaches such as SuRE can test largely every heterozygote variant in the genome
440 irrespective of its population-wise allele frequency or distance to known genes (van
441 Arensbergen et al., 2019). Likewise, for most variants the confounder of LD is removed,
442 as each variant is tested independently, allowing for the effect of individual variants on
443 transcription to be characterised. MPRA approaches do though have their own limitations,
444 most notably the fact that the variants are tested outside of their native sequence and
445 chromatin environment. Likewise, they can only be carried out in suitably transfectable
446 cell lines. These methods are therefore complementary approaches for defining
447 regulatory variation.
448
449 A further advantage of examining SuRE data is that the data is potentially more
450 comparable between species, having been generated using the exact same approach
451 (SURE), though in different cell types. This allowed us to further explore whether
452 differences in model performances between species was due to differences in the training
453 sets of regulatory variants. Consequently, in addition to using regulatory variants from the
454 GTEx project, we also built the models using regulatory variants (raQTLs) based on the
455 SuRE technique. For our human SuRE models, we included regulatory variants identified
456 in two cell lines: K562 (a myelogenous leukemia cell line) and HepG2 (a liver
457 hepatocellular carcinoma cell line) (van Arensbergen et al., 2019). We also trained a
458 model on the combined dataset of these two lines. These three human SuRE models
459 achieved test AUROC scores of 0.82 (HepG2), 0.84 (K562), and 0.81 (combined dataset)
460 on their respective test sets. Their ROC curves and cross-validation results are shown in
461 Supplementary Fig. 11. Additionally, we identified the top 30 most important features in
462 raQTL models (Supplementary Fig. 12). Among all the features, distance to motifs,
463 distance to cell type-specific chromatin data, and conservation score were consistently
464 ranked among the top features in both cell lines. While features such as distance to TSS
465 and gene density, which are among the top features in the GTEx models, did not appear
466 in the top-ranked features of raQTL models. This difference likely reflects the distinct
467 regulatory landscapes captured by these models. SuRE assays measure the intrinsic

regulatory potential of DNA sequences in a standardized context and regulatory environment independent manner. As a result, features related to intrinsic sequence properties, such as distance to motifs and conservation scores, rank higher in raQTL models.

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, broadly comparable to that observed for the human models. 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. We compared the top 30 most important features in the cattle raQTL model with the human raQTL models and found that distance to motifs and conservation scores ranked among the top features in both species, as shown in Supplementary Fig 12. Consequently, the difference between the human and cattle models is not apparent here when trained on more comparable datasets of regulatory variants.

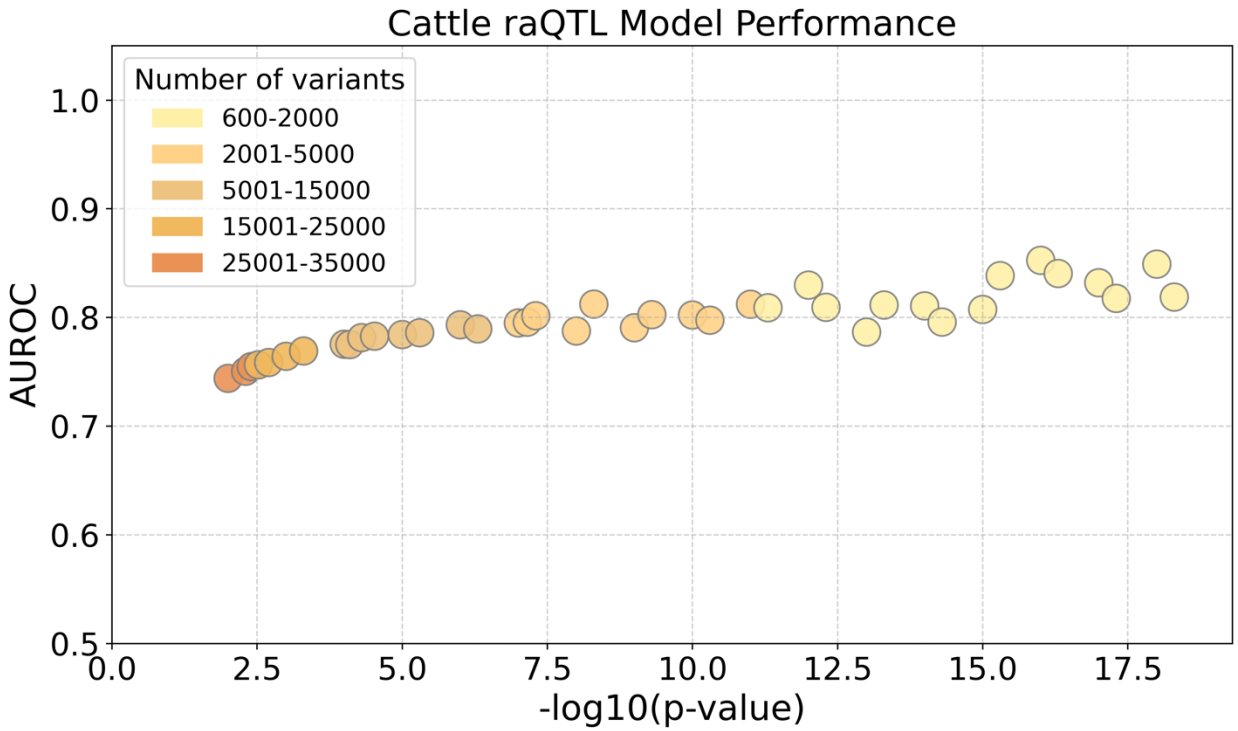


Figure 7 Cattle raQTL model performance when training and testing on regulatory variants selected based on different FDR adjusted p-value thresholds. For better clarity, the X axis is transformed to $-log_{10}^{p-value}$. The colour of each point represents the number of raQTL variants above the given 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 75% 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 2.3 compared to a background dataset of the same size, with randomly selected variants (repeated 10 times), where an average of 32% were predicted as regulatory at the same threshold (standard error of the mean: 0.62%).

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 recalculated. As shown in Figure 8B, it is possible to observe over 10-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, 58.0% of these validated regulatory variants were predicted as regulatory, compared to an average of 22.9% of matching background variants randomly selected from the 1000 Genomes cohorts (repeated 10 times, standard error of the mean: 0.71%). 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, with ratios of up to 9-fold observed. 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.

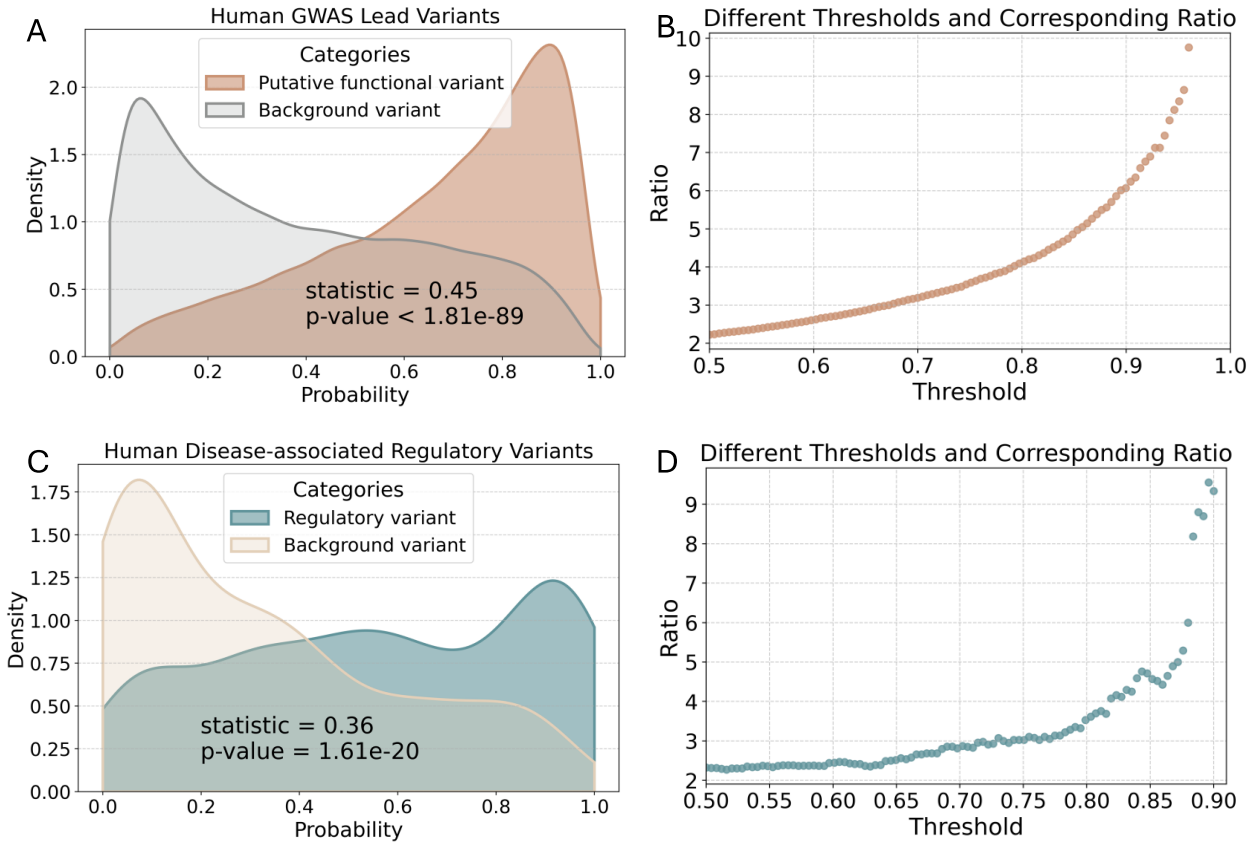


Figure 8 Prioritising human functional variants (A) Distributions of model derived prediction probabilities of putative human GWAS functional variants and background variants. The Kolmogorov-Smirnov statistic and p-value are shown in the plot. (B) Different prediction probability thresholds and their 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 four cattle traits: body weight, stature, body condition score (Reynolds et al., 2021), and milk yield (Jiang et al., 2019). Using GCTA-COJO, we selected independently associated SNPs for body weight, stature, and body condition score based on the corresponding GWAS summary statistics data. Specifically, 344, 241, and 96 fine-mapped SNPs were selected from the conditional analyses respectively. Additionally, we obtained 210 fine-mapped variants for milk yield from Jiang

546 et al.. Table 2 summarizes the proportion of these variants predicted as regulatory using
547 our different cattle models. All models predicted a higher proportion of fine-mapped
548 variants as regulatory compared to the equivalent number of matching background
549 variants. At the default prediction probability threshold of 0.5, the ratios of the proportion
550 of fine-mapped GWAS variants predicted as regulatory to the proportion of background
551 variants predicted as regulatory ranged from 1.5 to 3.0. Figure 9 A,B,C illustrate the
552 prediction probability distribution differences between fine-mapped and background data
553 in the three traits, body weight, body condition score, and milk yield. The prediction
554 probability distribution plots for stature can be found in Supplementary Fig. 13. We further
555 explored the impact of different prediction probability thresholds in the cross-species
556 model on the ratios of the proportion of fine-mapped variants predicted as regulatory to
557 background proportion. As shown in Figure 9D and E, 10-fold and 18-fold enrichment
558 ratios are observed at higher thresholds for body condition and milk yield, respectively.
559 This suggests that a higher prediction probability threshold could be useful for identifying
560 high-confidence functional variants in GWAS datasets (at the cost of a higher false
561 negative rate).
562
563 We identified variants that were consistently predicted as regulatory across different cattle
564 models. Among the 96 body condition score variants, 24 of them were predicted as
565 regulatory by all six models with a prediction probability exceeding 0.5. Similarly, 74 out
566 of 344 body weight-associated SNPs, 65 out of 241 stature-associated SNPs, and 77 out
567 of 210 milk yield-associated SNPs were consistently predicted as regulatory by all six
568 models with a prediction probability greater than 0.5 (Table 2). Supplementary Fig. 14
569 illustrates three examples. The chr14:23070083_C/T SNP, located near the TSS region
570 of the *TGS1* gene, was predicted as a regulatory variant associated with stature
571 (Supplementary Fig. 14A). This gene has been found to be associated with stature in both
572 human and cattle populations (Zhao et al., 2015). The body weight-associated variant
573 chr5:103942835_G/A is found in the intergenic region upstream of the *CD27* gene
574 (Supplementary Fig. 14B), which has previously been associated with weight gain in mice
575 (Englebert et al., 2024). Another predicted body condition score associated variant,
576 chr2:125968341_G/A, was found near the *WDTC1* gene (Supplementary Fig. 14C). This

gene has been associated with fat accumulation and obesity in humans (Ducos et al., 2017).

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

			Model type												
			Cross-tissues		**Adipose**		**Blood**		**Liver**		**Mammary**		**Muscle**		**Consistently regulatory**
Cattle Traits	**Body condition score (96[a])**	foreground	34 **(35.4%)**	**2.0**	25 **(26.0%)**	**1.5**	26 **(27.1%)**	**1.6**	33 **(34.4%)**	**1.8**	27 **(28.1%)**	**1.6**	27 **(28.1%)**	**1.6**	**24 (25.0%)**
		background	17±0.31 **(17.7%)**		17±0.37 **(17.7%)**		16±0.34 **(16.7%)**		18±0.28 **(18.8%)**		17±0.38 **(17.7%)**		17±0.37 **(17.7%)**		
	Body weight (344[b])	foreground	109 **(31.7%)**	**1.9**	107 **(31.1%)**	**2.0**	95 **(27.6%)**	**1.8**	96 **(27.9%)**	**1.6**	108 **(31.4%)**	**1.8**	110 **(32.0%)**	**1.9**	**74 (21.5%)**
		background	58±0.61 **(16.9%)**		54±0.65 **(15.7%)**		53±0.57 **(15.4%)**		60±0.72 **(17.4%)**		60±0.69 **(17.4%)**		59±0.68 **(17.2%)**		
	Stature (241[c])	foreground	82 **(34.0%)**	**2.0**	84 **(34.9%)**	**2.2**	78 **(32.4%)**	**2.1**	86 **(35.7%)**	**2.0**	85 **(35.3%)**	**2.0**	87 **(36.1%)**	**2.1**	**65 (27.0%)**
		background	41±0.58 **(17.0%)**		38±0.47 **(15.8%)**		38±0.54 **(15.8%)**		44±0.62 **(18.3%)**		43±0.56 **(17.8%)**		42±0.60 **(17.4%)**		
	Milk yield (210 [d])	foreground	106 **(50.5%)**	**2.7**	90 **(42.9%)**	**2.4**	95 **(45.2%)**	**2.7**	121 **(57.6%)**	**3.0**	95 **(45.2%)**	**2.4**	105 **(50.0%)**	**2.6**	**77 (36.7%)**
		background	39±0.50 **(18.6%)**		37±0.52 **(17.6%)**		35±0.50 **(16.7%)**		40±0.55 **(19.0%)**		40±0.56 **(19.0%)**		40±0.54 **(19.0%)**		

[a,b,c,d] Numbers of fine-mapped GWAS variants in body condition score, body weight, stature, and milk yield respectively. Equivalent numbers of variants were selected as background variants in each trait for comparison. Background variants were down-sampled 100 times to match the number of foreground variants, and the values reported represent the average number of background variants predicted as foreground, along with the standard error (SE).

[*] 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 blue cells in the last column indicate the count of variants consistently predicted as regulatory by all 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.

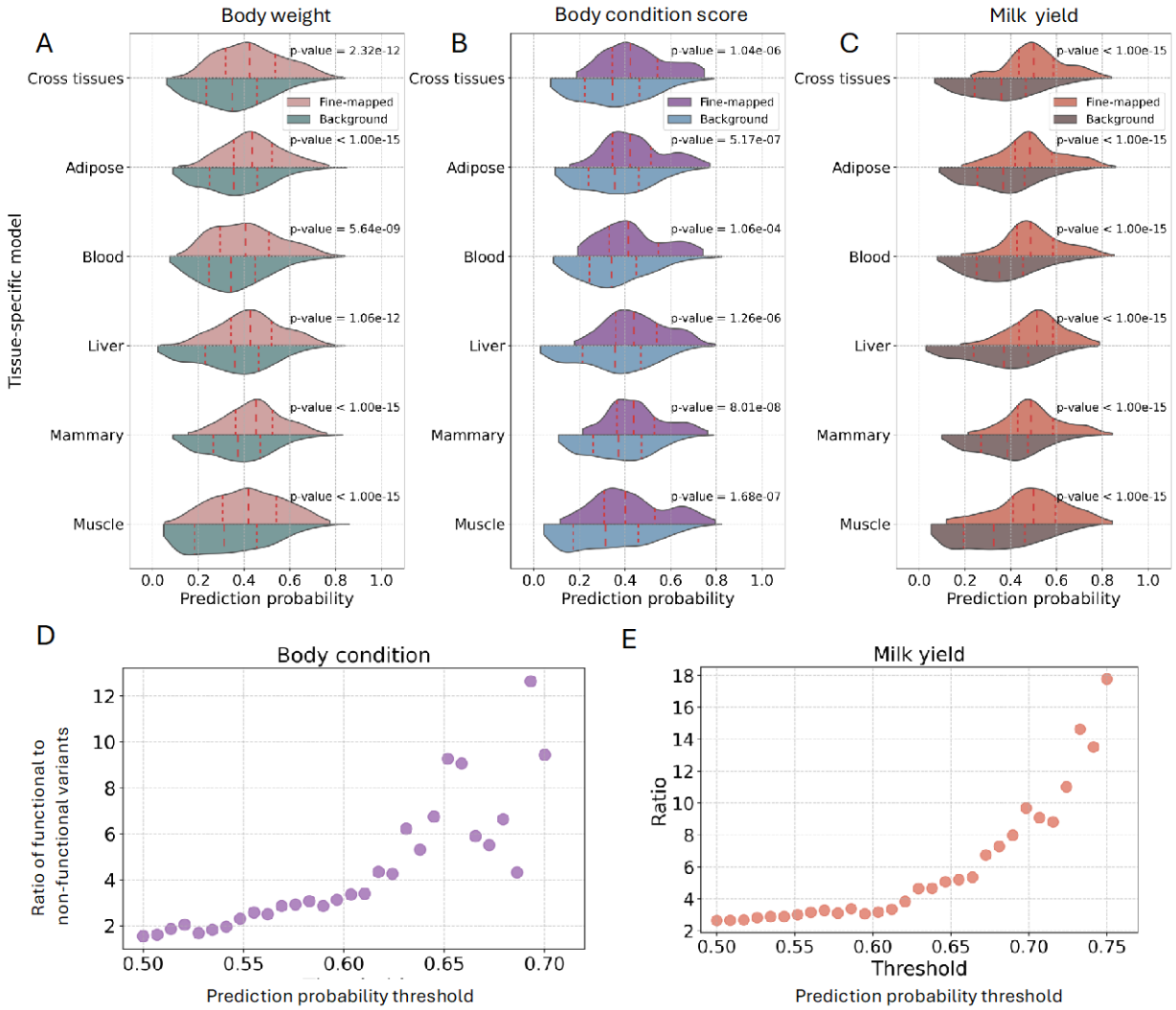


Figure 9 Prioritising cattle functional variants. Distributions of model derived prediction probabilities of cattle GWAS fine-mapped and background variants for (A) body weight, (B) body condition, and (C) milk yield. The red dashed lines in the violin plots represent the median and quartile values. 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 body condition score predicted as regulatory to the proportion of background variants predicted as regulatory using the cross-tissues model. (E) Different prediction probability thresholds and their corresponding ratios of the proportion of cattle GWAS lead variants associated with milk yield predicted as regulatory to the proportion of background variants predicted as regulatory using the cross-tissues 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 when trained just on the baseline features marked in the salmon colour in Table 1. 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 cattle models may have possible potential for prioritising regulatory variants and elements linked to important cattle traits that could be explored in further work.

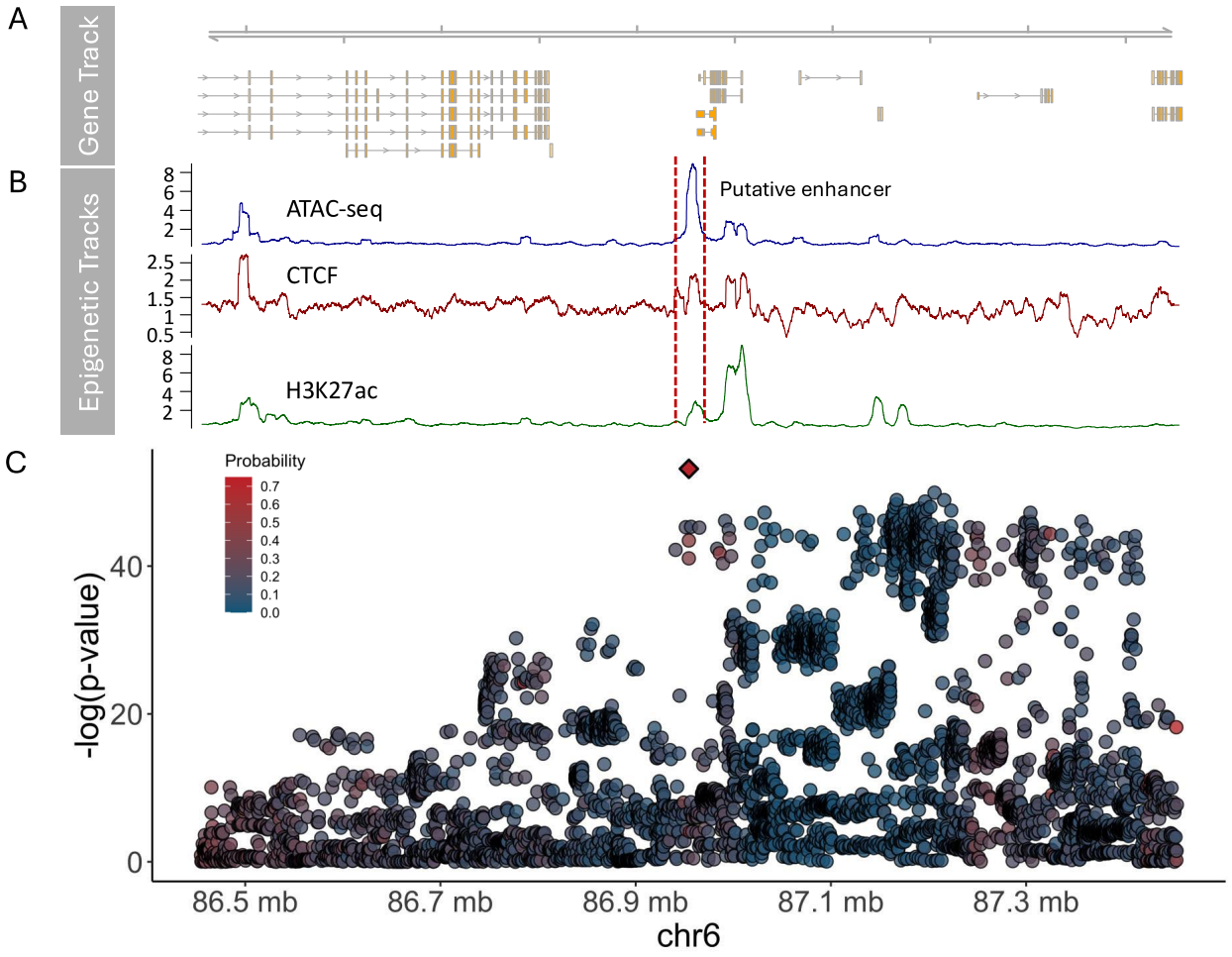


Figure 10 Regulatory predictions at the GC locus linked to traits including mastitis resistance and body condition score. (A) The genes and different isoforms at this locus. (B) ATAC-seq (Lee et al., 2021), CTCF and H3K27ac (Kern et al., 2021) tracks in this region, with the location of the putative 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 remains a considerable challenge, in large part because they typically occur in less well-annotated non-coding regions and are affected by linkage disequilibrium (LD) among neighbouring variants. These problems are even more pronounced in livestock species, where high-quality functional genomic data is relatively scarce and LD tends to be extensive. To address this challenge, we developed

633 a species-agnostic variant annotation pipeline that integrates a range of features, from
634 sequence conservation scores to positional properties, and applied it to build human and
635 cattle machine learning models for predicting variants with regulatory effects.
636

637 Our models, trained in both humans and cattle, effectively distinguished regulatory
638 variants from non-regulatory variants, and similar predictive features emerged as
639 important in both species. Earlier work on prioritising cattle functional variants had
640 highlighted the particular value of conservation metrics (Xiang et al., 2019). However, that
641 study had relied on lifting over human conservation metrics to the cattle genome which is
642 problematic because only about half of human bases can be directly mapped to a
643 corresponding cattle base. In the present study, we generated PhastCons and PhyloP
644 conservation metrics specifically for cattle for the first time, confirming their importance
645 for predicting regulatory variants in line with previous findings. Although the multiple
646 alignment underlying these scores was based on the Btau_5.0.1 reference genome, 98.7%
647 of the metric sites successfully lifted over to BosTau9, illustrating that the new cattle-
648 specific conservation metrics can be effectively utilised with current genome builds.
649

650 The performance of both the human and cattle models exhibited a slight downward trend
651 in performance with increasing distance from TSSs. This pattern is perhaps expected
652 given the more defined regulatory code near promoters and other well-characterized
653 transcriptional control regions. However, despite this modest decline, the overall stability
654 of performance across all distance bins suggests that our models have the ability to
655 generalize across the genome, including regions far from canonical regulatory hubs.
656 Interestingly the NRF1 motif was found to be the most informative motif for predicting
657 regulatory variants and the seventh most important feature overall in both species when
658 training was restricted to the set of common features. This may result from NRF1 being a
659 ubiquitously expressed canonical activator with a relatively well defined motif (Duttke et
660 al., 2024).
661

662 Applying our GTEx-based models to human and cattle datasets demonstrates their
663 effectiveness in prioritizing functional variants associated with important traits. While the

664 precise circumstances and full extent of improvements remain a topic of ongoing research,
665 previous studies suggest that incorporating functional variant-based SNP panels may
666 enhance trait prediction accuracy (Santana et al., 2023; Xiang et al., 2021). Consequently
667 the enrichments of up to 18-fold for trait-associated variants holds potential for the
668 prioritization of variants for calculating genomic estimated breeding values (gEBVs) in
669 livestock. Unlike the requirement for the direct identification of functional variants to
670 improve livestock using approaches such as genome editing, the identification of causal
671 variants is not always required when prioritizing SNPs for gEBV calculation, provided that
672 the variants in LD with the true causal alleles are enriched for (Xiang et al., 2021).
673 Consequently, perfect classification accuracy may not be strictly necessary to achieve
674 practical gains. Simply enriching the SNP panel used for gEBV calculations with
675 functional variants, or tagging variants in high LD, has the possibility to improve their
676 accuracies, and potentially lead to substantial cumulative benefits over time.
677
678 The analysis of the GC locus also highlights the possible potential of these models for
679 fine-mapping functional regulatory elements at GWAS loci. Despite not being trained on
680 the GWAS or chromatin data, the model identified the lead GWAS variant at the
681 previously identified enhancer element as the variant with the highest regulatory potential.
682 So, although the functional variant at this locus is thought to be a structural variant, this
683 result highlights how relevant regulatory elements may potentially be prioritised using
684 these models for further follow up study.
685
686 To date, deep learning-based models, especially those popular models in natural
687 language processing (NLP), have been explored for understanding DNA sequences.
688 However, the prediction results in our study illustrate that Enformer features did not
689 provide additional substantial discriminative information beyond our general features for
690 both human and cattle models. This aligns with recent studies evaluating popular deep
691 learning architectures, including Enformer, for predicting gene expression variation
692 (Huang et al., 2023), as well as *in silico* investigations focusing on Enformer's ability to
693 capture expression in promoters and enhancers (Karollus et al., 2023). These
694 observations reported limitations in these models' performance. Specifically, they

695 observed limited effectiveness in elucidating the impact of variants on gene expression
696 and difficulties in capturing the causal influence of distal enhancers on gene expression.
697 Although these previous studies primarily focused on Enformer's ability to predict gene
698 expression levels, rather than simply define the location of regulatory variants, these
699 observations regarding Enformer's limitations align with the results of our study,
700 highlighting the challenges deep learning models face in capturing the complex regulatory
701 effects of individual base changes.
702
703 The difference in performance between human and cattle models, even when using
704 common features, suggests limitations in the cattle variant data. This is further supported
705 by the minimal improvement (~0.03 AUROC score increase) observed in human models
706 trained on all general features compared to models using only common human-cattle
707 features, and by the fact that adding 95 cattle chromatin features did not enhance the
708 performance of the cattle model. The comparatively small gap in performance of the cattle
709 and human models trained on the more comparable MPRA data also supports the idea
710 that the difference in performance of the cattle and human eQTL models is because the
711 foreground sets are less well matched, i.e. the core issue lies in the lack of a gold-
712 standard regulatory variant set for training cattle models, which is largely responsible for
713 their relatively low performance. This limitation is likely exacerbated by strong linkage
714 disequilibrium (LD) in cattle populations, which complicates fine-mapping efforts from
715 eQTL data and may introduce uncertainty in regulatory variant identification. While the
716 cattle GTEx project serves as one of the most extensive reference resources for the cattle
717 transcriptome available to date, it does have challenges. This includes a comparatively
718 limited number of individuals and breeds for certain cell types, the dependence on calling
719 variants from the RNA-seq data, and difficulty of accurate variant imputation in
720 underrepresented breeds. Additionally, unlike human GTEx, cattle GTEx data comes
721 from diverse studies, potentially reducing its reliability due to variable batch effects.
722 Together these factors may be leading to the difference in model performance between
723 human and cattle. Addressing these challenges requires larger sample sizes in consistent
724 cohorts, matching whole-genome variant calls, cross-breed analyses to break down
725 breed-specific LD patterns, and potentially improved functional annotation efforts using

high-throughput cellular screens. A more comprehensive collection of high-quality cattle data, as well as data for other species, encompassing high quality functional data 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,

757 the conservation scores for cattle were computed using the phastCons and phyloP tools
758 (Hubisz et al., 2011) with the neutral model serving as the phylogenetic model. Since the
759 241-way mammalian alignment employed the Btau_5.0.1 cattle genome assembly, we
760 conducted a lift-over step to map the conservation scores to the updated BosTau9
761 assembly (Rosen et al., 2020). Due to the lack of lift-over chain file between these two
762 assemblies, nf-LO (Talenti & Prendergast, 2021) was used to generate the chain file.
763 Subsequently, the USCS liftOver tool was used to perform the mapping, successfully
764 aligning 98.7% of variants to the BosTau9 cattle assembly.
765
766 For the variant position properties, distances from the variants to various genomic
767 elements were calculated. The genomic locations of CpG islands for different species
768 were downloaded from UCSC (Nassar et al., 2023). A total of 1,554 distinct types of
769 human chromatin data were obtained from Ensembl (Martin et al., 2023). For cattle, 95
770 chromatin data features were obtained (Kern et al., 2021). Distances from the variants to
771 the nearest CpG island and chromatin region were calculated using the bedtools closest
772 function (Quinlan & Hall, 2010). The R library AnnotationHub was used to obtain
773 transcription start sites (TSS) for different species, and only biotypes with greater than
774 1000 TSS were retained. The human regulatory features, including enhancers,
775 transcription factors (TF), and CTCF binding sites, were obtained from Ensembl using the
776 R library biomaRt (Smedley et al., 2009). Distances to the nearest transcription start site
777 (TSS) and regulatory features (human only) were calculated using the ChIPpeakAnno R
778 library (Zhu et al., 2010). Gene density per megabase pair (1Mb) for different species was
779 calculated using the AnnotationHub R library (Gentleman et al., 2004) and
780 GenomicRanges library (Lawrence et al., 2013). The VEP command line tool (McLaren
781 et al., 2016) was used to obtain variant consequences. To capture the sequence context
782 of the variants, except the allele change, we extracted the 5-mer flanking sequence
783 centered on the target variant. Reference genomes for the two species in FASTA format
784 were obtained from UCSC (Nassar et al., 2023) and the Samtools faidx command was
785 used to extract the flanking sequence from the reference genome (Li et al., 2009). To
786 calculate the distances from variants to different motifs, scanMotifGenomeWide.pl in
787 HOMER (Duttke et al., 2024) was used to predict motif sites across both the human and

788 cattle genomes. The 460 known vertebrate motifs provided by HOMER were used as
789 input for genome-wide scanning. After obtaining the genome-wide positions of the motifs,
790 bedtools closest (Quinlan & Hall, 2010) was used to calculate distances.
791
792 In addition to general annotations, 5313 predicted functional genomic data for each
793 variant were obtained using the Enformer model (Avsec et al., 2021). This includes a
794 diverse set of functional genomic signals, including chromatin accessibility (ATAC-seq,
795 DNase-seq), histone modifications (ChIP-seq), transcription factor binding sites, and
796 gene expression (CAGE-seq). The pre-trained Enformer model was loaded from
797 Tensorflow hub (https://tfhub.dev/deepmind/enformer/1), and the reference genome in
798 FASTA format for the target species was downloaded from UCSC (Nassar et al., 2023).
799 Python code for running the pre-trained Enformer model was adopted from the Enformer
800 GitHub page (https://github.com/deepmind/deepmind-research/tree/master/enformer).
801 The pre-trained model was employed to make predictions for the sequences (393,216 bp)
802 centered on the reference and alternative alleles, respectively, where 393,216 bp
803 represents the sequence window used as input to Enformer. Subsequently, the 5313-
804 column scores for each variant were defined as the differences between the reference
805 and alternative prediction results.
806
807 To facilitate the reusability of the aforementioned annotation approaches, a SNV
808 annotation pipeline was developed using the Nextflow framework (Di Tommaso et al.,
809 2017). These annotation approaches were encapsulated into separate processes, which
810 can run in parallel to enhance annotation efficiency. They were organized into two sub-
811 workflows, one for general annotations and another for Enformer scores. The default
812 values of parameters and profiles for various execution environments were defined in the
813 configuration files. The pipeline can be accessed from: https://github.com/evotools/nf-
814 VarAnno.
815
816 **Genotype datasets**
817 Various variant sets were used in the study to explore the utility of annotations in
818 regulatory variant modelling. Human fine-mapped regulatory variants were obtained from

819 the CaVEMaN dataset from the Genotype-Tissue Expression (GTEx) project (GTEx
820 Consortium, 2020). We focused on autosomal SNPs, resulting in a set of 590,778
821 regulatory variants (foreground data) for analysis. For obtaining high-confidence
822 regulatory variants, we filtered the 590,778 variants to those with a CaVEMaN causal
823 probability > 0.5, resulting in 45,987 variants, where CaVEMaN is a statistical framework
824 designed to prioritize causal variants (Brown et al., 2017). To maintain consistency with
825 the regulatory variants, matching sets of background variants were down-sampled from
826 variants located within 1Mb of the TSS and selected to match the minor allele frequency
827 (MAF) distribution of the regulatory variants. The MAF of human variants were obtained
828 from VEP (McLaren et al., 2016a). Additionally, to benchmark the performance of our
829 approach, we obtained 11.4k variants, which include 1,140 putative functional regulatory
830 variants across 83 human complex traits and 10,260 control variants, used in a recent
831 paper benchmarking models for causal regulatory variant prediction in human (Benegas,
832 Eraslan, et al., 2025).
833
834 Similarly, regulatory variants from cattle GTEx were acquired for cattle modelling (Liu et
835 al., 2022). Lead variants, denoting those with the smallest nominal p-value for each gene
836 across all tissues, were extracted. Furthermore, these variants were filtered based on
837 their associated permutation p-values, using a threshold of 0.05. After filtering, 79,215
838 SNPs were retained as regulatory. Matching sets of background variants were down-
839 sampled from the remaining dataset, selecting only variants with a nominal p-value
840 greater than 0.05 and ensuring that their MAF distribution matched that of the foreground
841 variants. Cattle MAF was calculated using BCFtools based on the reference panel from
842 cattle GTEx (Liu et al., 2022).
843
844 For both humans and cattle, we obtained another set of regulatory variants (raQTL) based
845 on the survey of regulatory elements (SuRE) technique. The human raQTLs were
846 obtained from van Arensbergen et.al. (van Arensbergen et al., 2019). The original variants
847 on hg19 were lifted over to hg38, resulting in 18,717 SNPs in the K562 cell line and 13,838
848 SNPs in the HepG2 cell line. The same number of background variants for these two cell
849 lines were randomly sampled from the tested variants with SuRE signals greater than 4

850 and p-values greater than the threshold for defining raQTLs in the original paper
851 (0.006192715 for K562 and 0.00173121 for HepG2). The cattle raQTLs were generated
852 following the same approach as in van Arensbergen et.al. and is described in full detail in
853 an accompanying paper. Instead of using the statistical approaches employed for
854 identifying raQTLs in humans, our group used zero-inflated negative binomial regression
855 models to analyze SuRE read count files. After filtering according to the adjusted p-value
856 with a threshold of 0.05, there were 66,501 variants kept for further analysis. The same
857 number of matching background variants were randomly sampled from the tested variants
858 with adjusted p-value greater than 0.05.
859
860 To assess the effectiveness of human and cattle GTEx models in predicting regulatory
861 effects of reported variants associated with traits, we used data from genome-wide
862 association studies. For human GWAS data, we used the variants from the credible sets
863 in Open Target Genetics (Ghoussaini et al., 2021). We downloaded the datasets from
864 Open Target Genetics and the Python package pyarrow (Richardson et al., 2023) was
865 employed to convert the data into a Python dataframe from parquet format. Afterward, the
866 variants were filtered based on a posterior probability threshold of 0.5, and those variants
867 also present in the human high-confidence training set were removed, resulting in 28,918
868 variants. For comparison purposes, an equal number of background variants were
869 randomly selected from the remaining variants with posterior probability smaller than 0.5.
870 Another set of pathogenic regulatory variants associated with Mendelian diseases was
871 obtained from Damina et al. (Smedley et al., 2016). The 406 single nucleotide variants
872 were initially lifted over from the GRCh37 assembly to the GRCh38 assembly using the
873 UCSC liftOver tool (Kuhn et al., 2013). After excluding those on sex chromosomes, 345
874 SNVs were retained for analysis. An equal number of background variants were randomly
875 selected from the human 1000 genomes cohort (1000 Genomes Project Consortium et
876 al., 2015).
877
878 For cattle GWAS data, we initially obtained the GWAS summary statistics data for three
879 cattle traits: body condition score, body weight, and stature, from Edwardo et al (Reynolds
880 et al., 2021). The originally reported variants on cattle genome UMD3.1 were lifted over

881 to ARS-UCD1.2 using the mapping file provided in the paper, successfully mapping 98.6%
882 of the variants between the assemblies. Subsequently we conducted conditional analysis
883 using GCTA-COJO to select independently associated SNPs for each trait (Yang et al.,
884 2012). To run GCTA-COJO, the initial summary statistics data were re-organized to match
885 the input format required by the command. PLINK 2.0 (Chang et al., 2015) was used to
886 generate the PLINK binary PED files, which served as an input for GCTA-COJO. To create
887 the reference VCF files needed for input into PLINK 2.0, we first downloaded the
888 reference VCF files from the 1000 Bull Genomes Project (Hayes & Daetwyler, 2019).
889 Then these files were filtered using bcftools (Danecek et al., 2021) to include only a subset
890 of samples that match the breeds included in the cattle GWAS data. After obtaining the
891 associated SNPs for these traits, they were filtered based on their p-values using a
892 threshold of 5e-8. Consequently, 96, 344 and 241 associated variants were retained for
893 body condition score, body weight, and stature, respectively. Additionally, we obtained
894 fine-mapping results for milk yield from Jiang et al. (Jiang et al., 2019). The variants from
895 the cattle genome UMD3.1 were lifted over to ARS-UCD1.2, the allele changes were
896 adjusted based on the liftOver chain file, and a p-value threshold of $5x10^{-8}$ was applied to
897 identify 210 associated variants for milk yield. For comparison, we obtained background
898 data with p-values greater than 0.05, selecting the same number of variants as the
899 foreground 100 times for each trait.
900
901 **Regulatory variant modelling**
902 The variants were annotated using the variant annotation pipeline. These annotations first
903 underwent feature engineering processes, performed using the Python scikit-learn
904 package (Pedregosa et al., 2011). The conservation score annotations contained missing
905 values, affecting approximately 0.9% of human variants and 0.8% of cattle variants.
906 Given the small proportion, we opted to exclude variants with missing conservation scores
907 from the dataset rather than employing imputation strategies. Subsequently, categorical
908 features were encoded using different strategies. For sequence categorical features, such
909 as 5-mer flanking sequence and allele change, we created a dictionary to map each
910 nucleotide in the sequence to a binary string (A: 1000, C: 0100, G: 0010, T: 0001). The
911 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.

Four machine learning algorithms were considered for model construction, Random Forest (Breiman, 2001), CatBoost (Prokhorenkova et al., 2017), XGBoost (T. Chen & Guestrin, 2016), and Support Vector Machines (SVM) (Cortes & Vapnik, 1995). Random Forest and SVM models were implemented using scikit-learn libraries, while CatBoost and XGBoost utilized their respective Python packages. Considering both training efficiency and model performance, CatBoost was selected as the algorithm for modelling in the study. Each variant set, including both foreground and background variants, was split into training and test sets by chromosomes to prevent data leakage. Specifically, several chromosomes were left out for testing, while the remaining chromosomes were used for model training and hyper-parameter tunning. To optimize hyper-parameters, we used BayesSearchCV from the Python scikit-optimize library (Head et al., 2018), a Bayesian optimization-based approach that efficiently tunes hyper-parameters through iterative cross-validation. For cross-validation, we implemented chromosome-level cross-validation, where entire chromosomes were held out in each fold to ensure that variants from the same chromosome did not appear in both training and validation sets. When performing benchmarking, we used a single-chromosome hold-out strategy for hyper-parameter tuning and final evaluation. 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. The model tuning and training were conducted using GPUs from the Edinburgh international data facility (EIDF).

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 (S. Lundberg & Lee, 2017).

Code availability

The cross-species variant annotation pipeline is available on GitHub: https://github.com/evotools/nf-VarAnno. The machine learning pipeline is available on GitHub: https://github.com/evotools/functional-variant-prediction.

Data availability

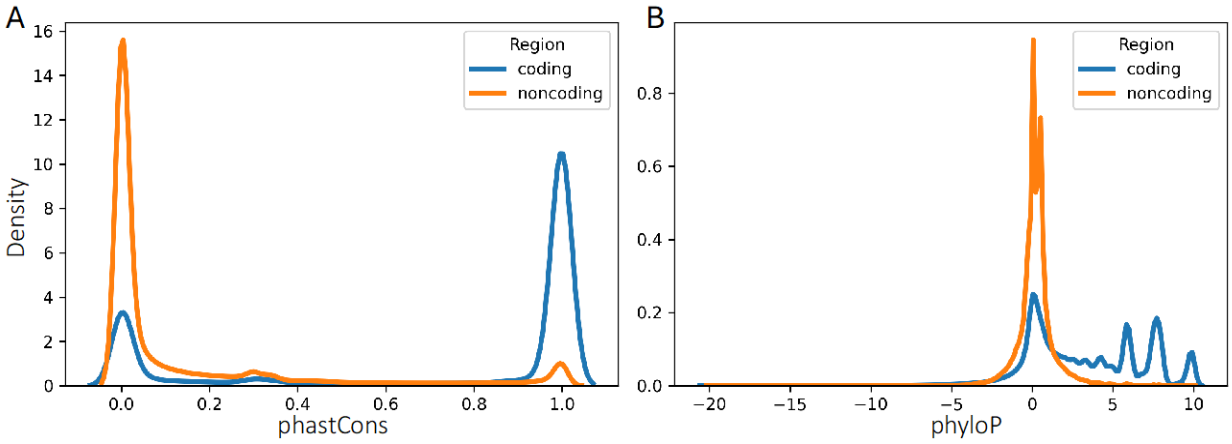
The cattle conservation metrics are available at: https://doi.org/10.5281/zenodo.13332541. The cattle tissue-specific and cross-tissue models for functional variant prediction can be found at: https://doi.org/10.5281/zenodo.14901001.

Acknowledgments

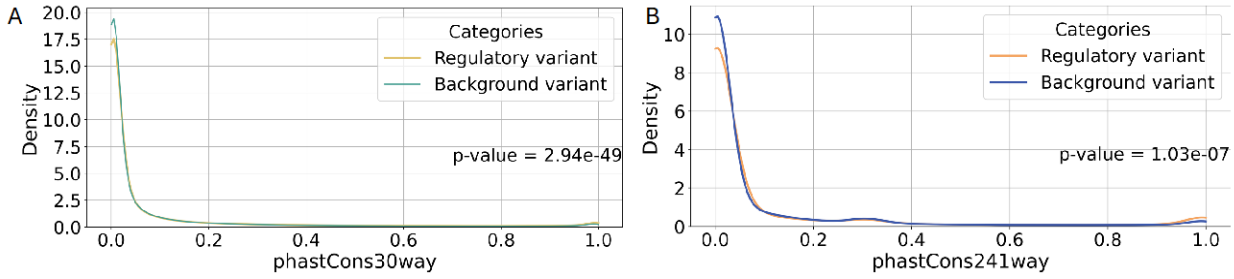
This work was supported by grant BB/W000288/1 and Institute Strategic Programme Grant BBS/E/D/10002070 from the Biotechnology and Biological Sciences Research Council (BBSRC). This project was also supported by funding by the Bill and Melinda Gates Foundation (BMGF) and with UK aid from the UK Government's 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. We would like to thank Dr Ivan Pocrnic for helpful discussions regarding this work.

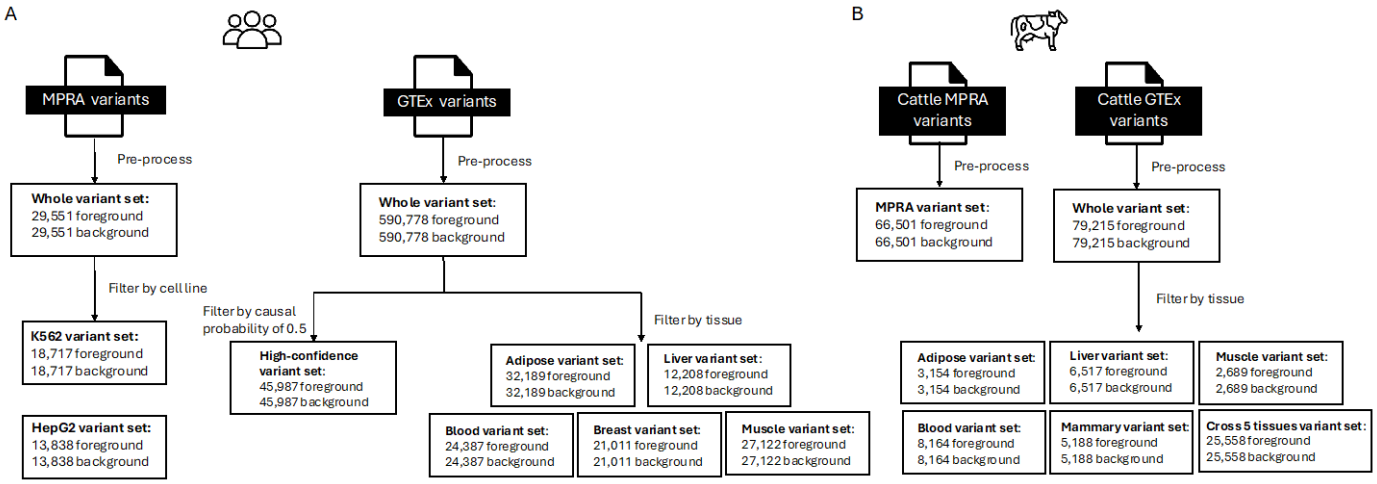
Supplementary materials



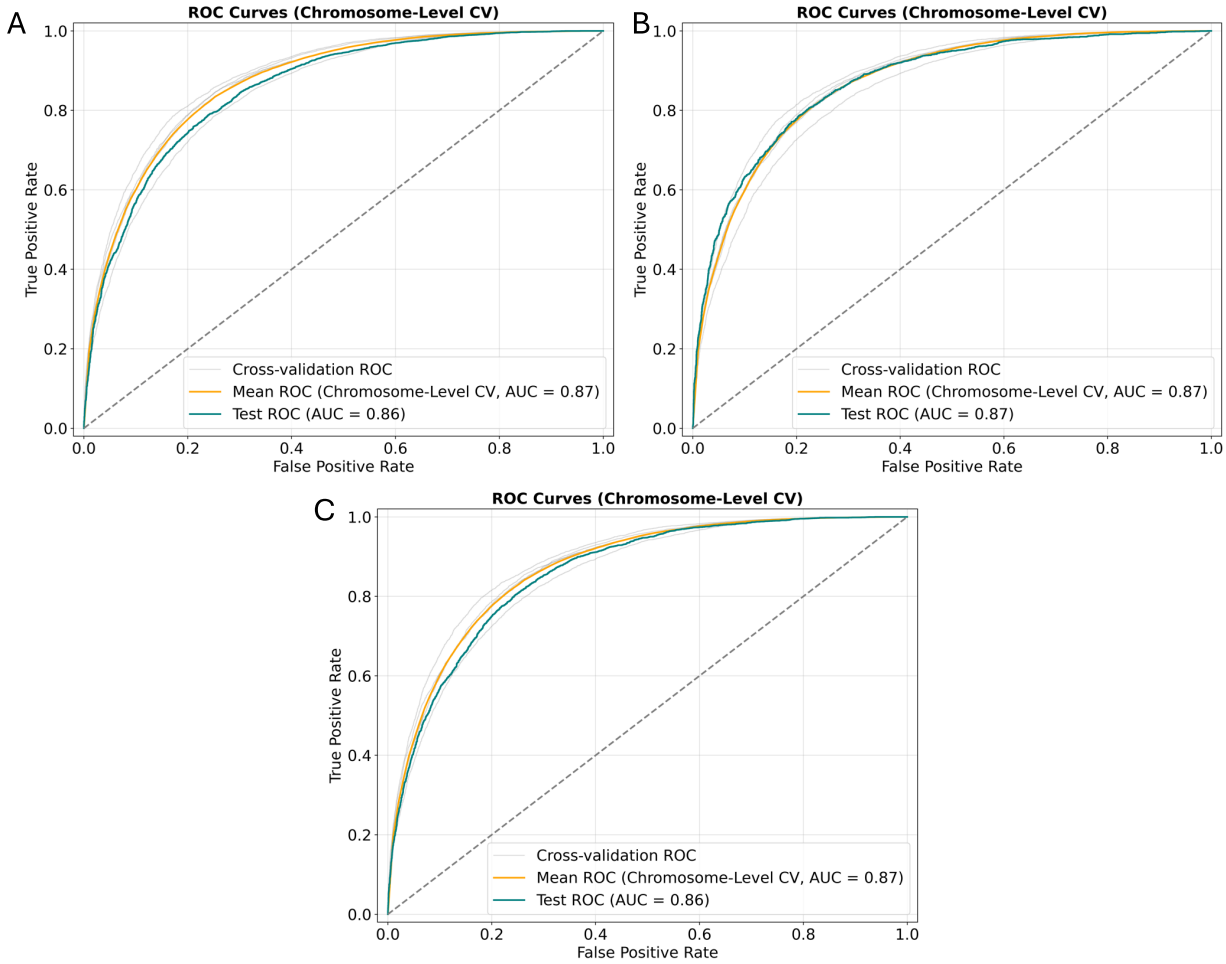
Supplementary Figure 1 Conservation score distribution differences in coding and noncoding regions in the cattle genome. (A) phastCons241way conservation score (B) phyloP241way conservation score.



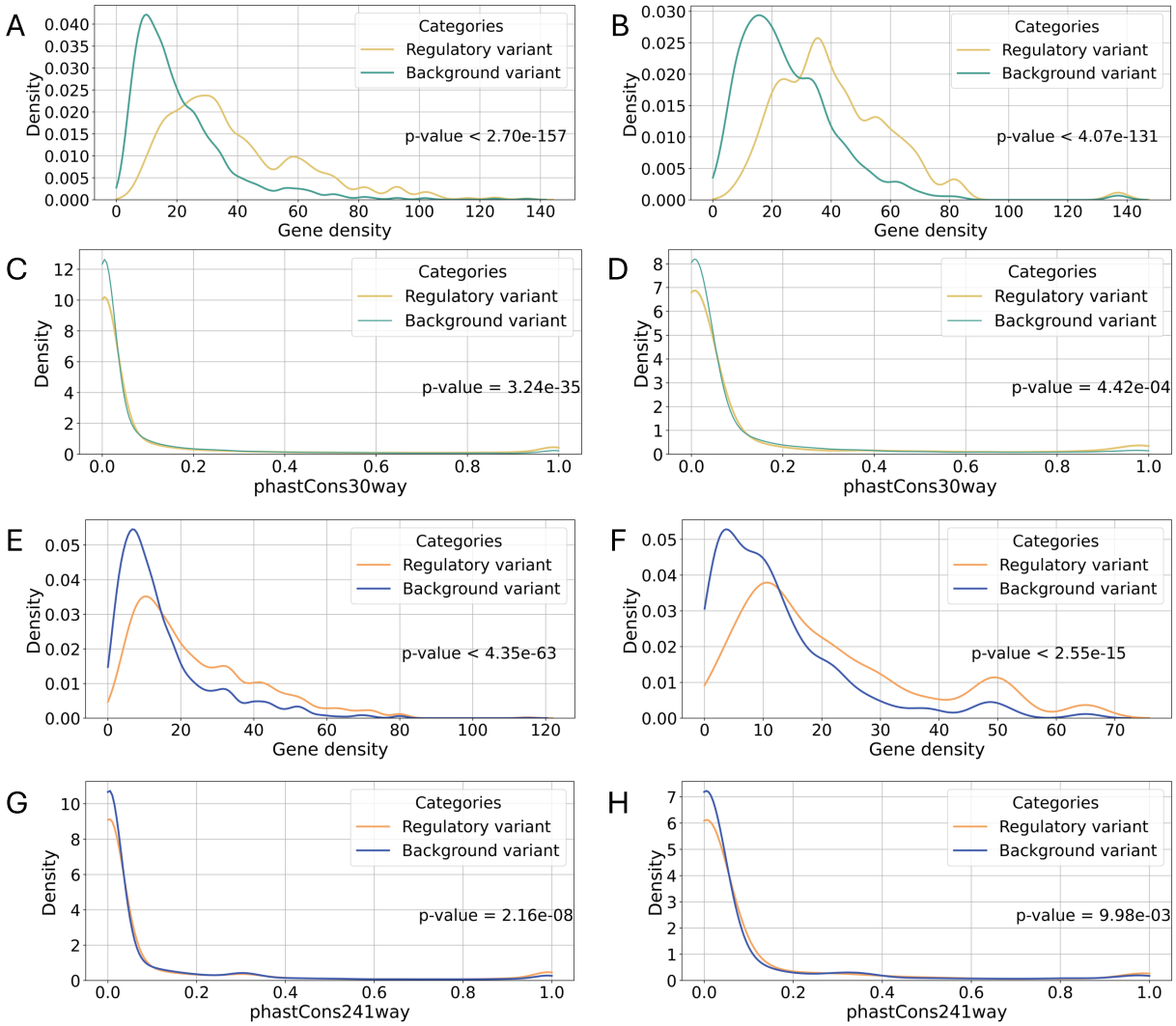
Supplementary Figure 2 Kernel density estimate (KDE) of the distribution of phastCons scores for (A) human and (B) cattle. The conservation scores display significant differences between the groups in both human and cattle (Two-sample Kolmogorov-Smirnov test p-values are shown in the corresponding plots).



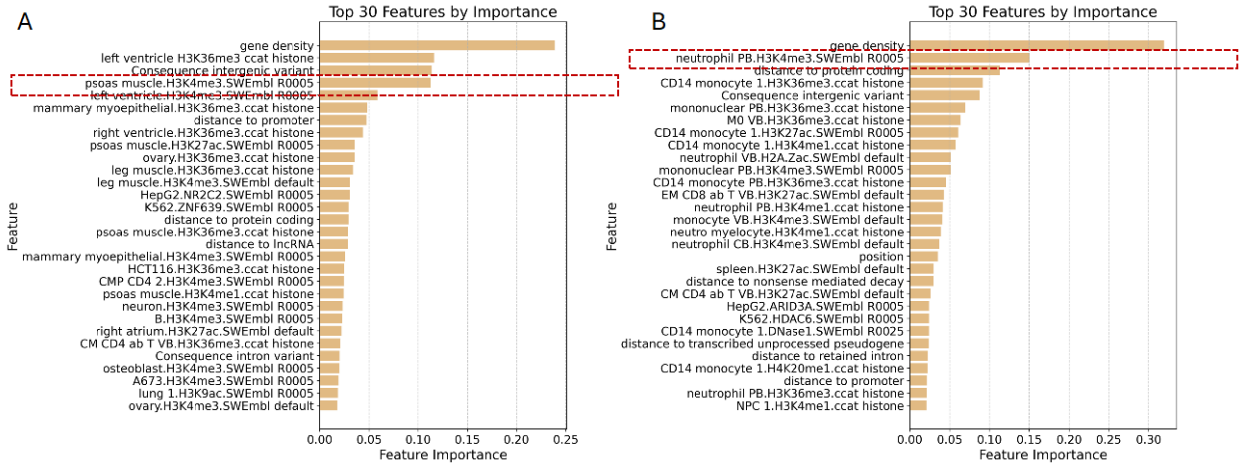
Supplementary Figure 3 An overview of different variant sets included in the study. (A) Human. (B) Cattle



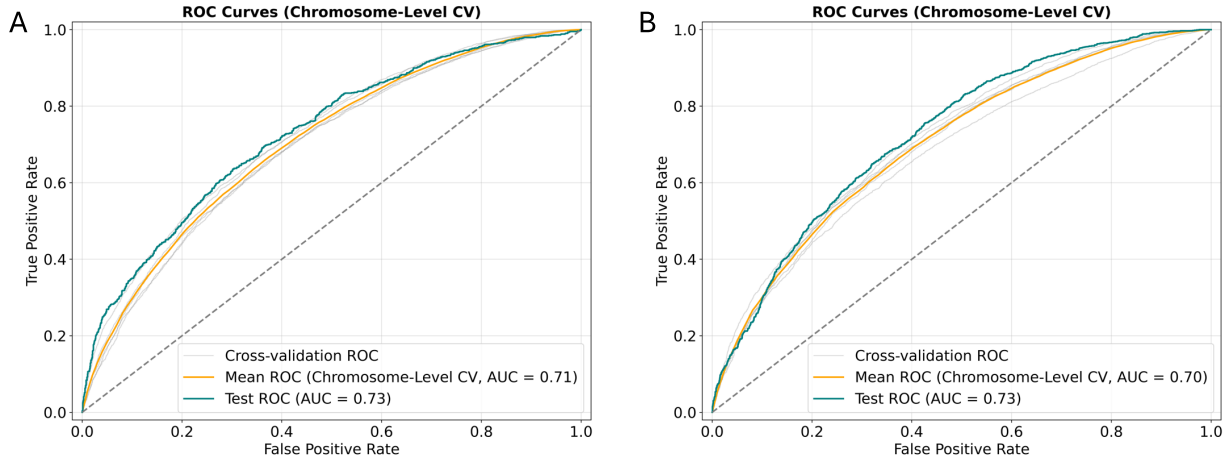
Supplementary Figure 4 ROC curves and AUROC (AUC) scores for the human high-confidence model when testing on (A) chr1 and chr22, (B) testing on chr3, and (C) testing on chr2 and chr21.



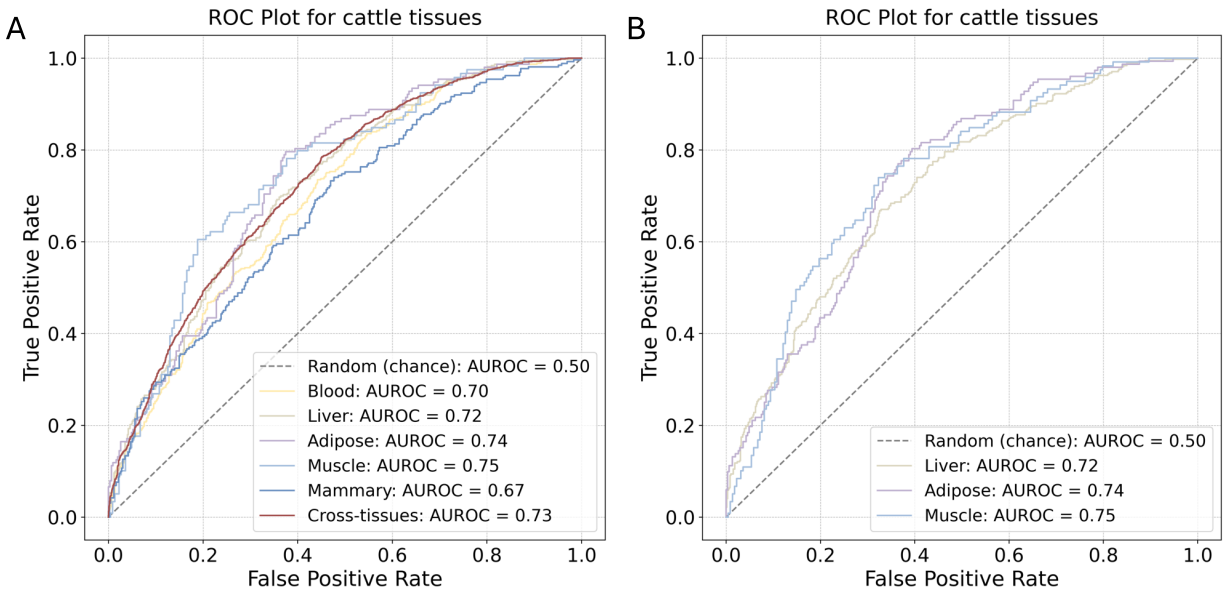
Supplementary Figure 5 Feature distributions of foreground and background variants in training and test sets. (A) Human gene density distribution in the training set. (B) Human gene density distribution in the test set. (C) Human phastCons30way distribution in the training set. (D) Human phastCons30way distribution in the test set. (E) Cattle gene density distribution in the training set. (F) Cattle gene density distribution in the test set. (G) Cattle phastCons241way distribution in the training set . (H) Cattle phastCons241way distribution in the test set. All distributions display significant differences between the groups in both human and cattle (Two-sample Kolmogorov-Smirnov test p-values are shown in the corresponding plots).



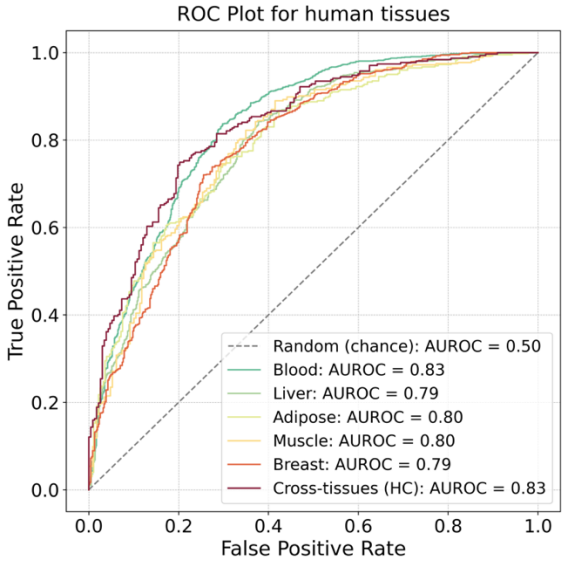
Supplementary Figure 6 Top 30 features in human (A) muscle and (B) blood specific models. The features are ranked in descending order according to their relative feature importance in the model. The features within the red dashed rectangle are the top chromatin features that exclusively emerged among the top 10 in their respective tissue-specific models.



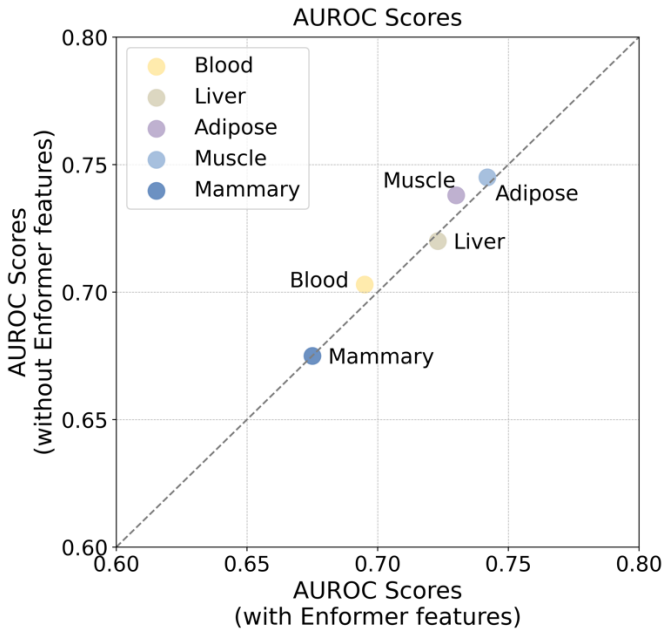
Supplementary Figure 7 ROC plots and AUROC (AUC) scores for cattle cross-tissues model when testing on (A) chr1 and chr28, (B) testing on chr10.



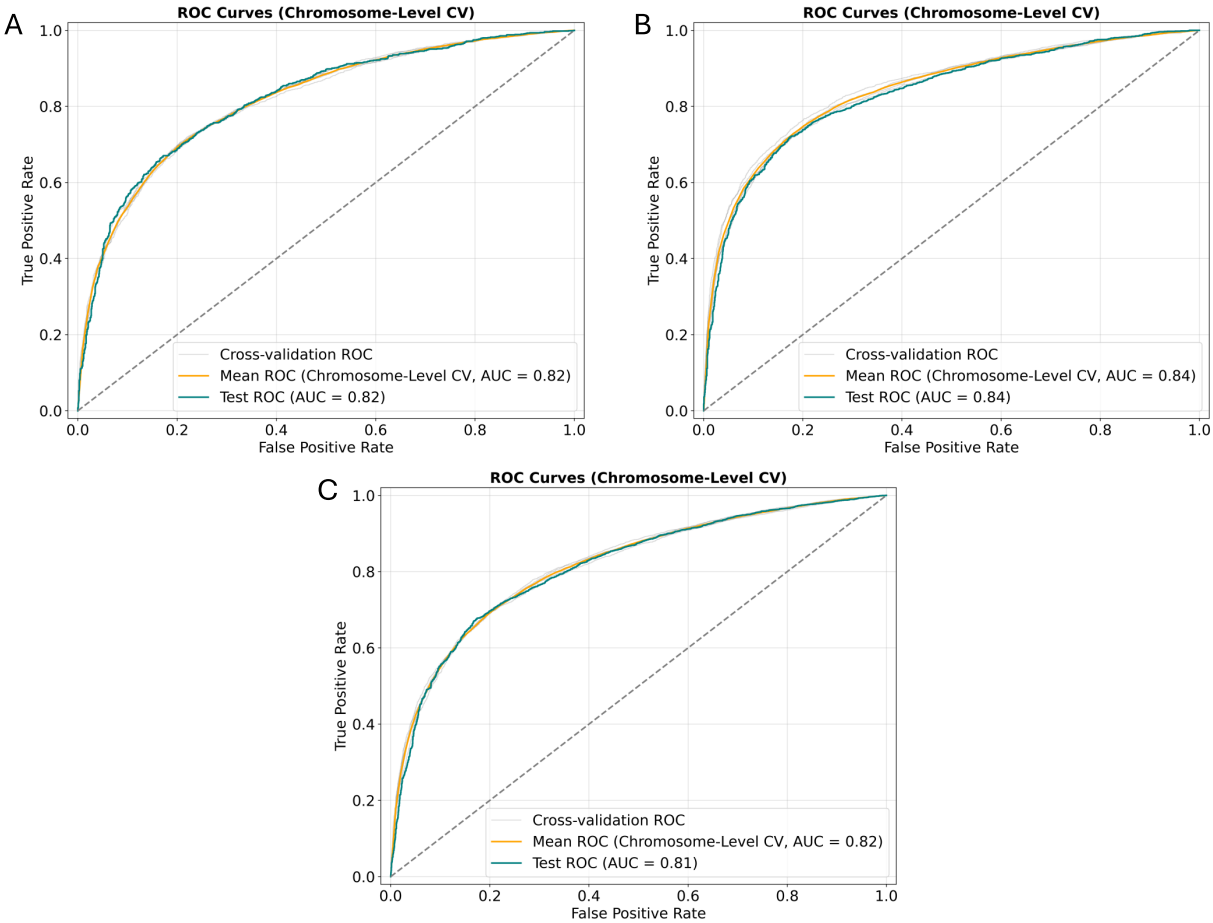
Supplementary Figure 8 ROC plots and AUROC scores for cattle model trained with (A) full set of 95 cattle chromatin data, and (B) with tissue-specific chromatin data for liver, adipose, and muscle models.



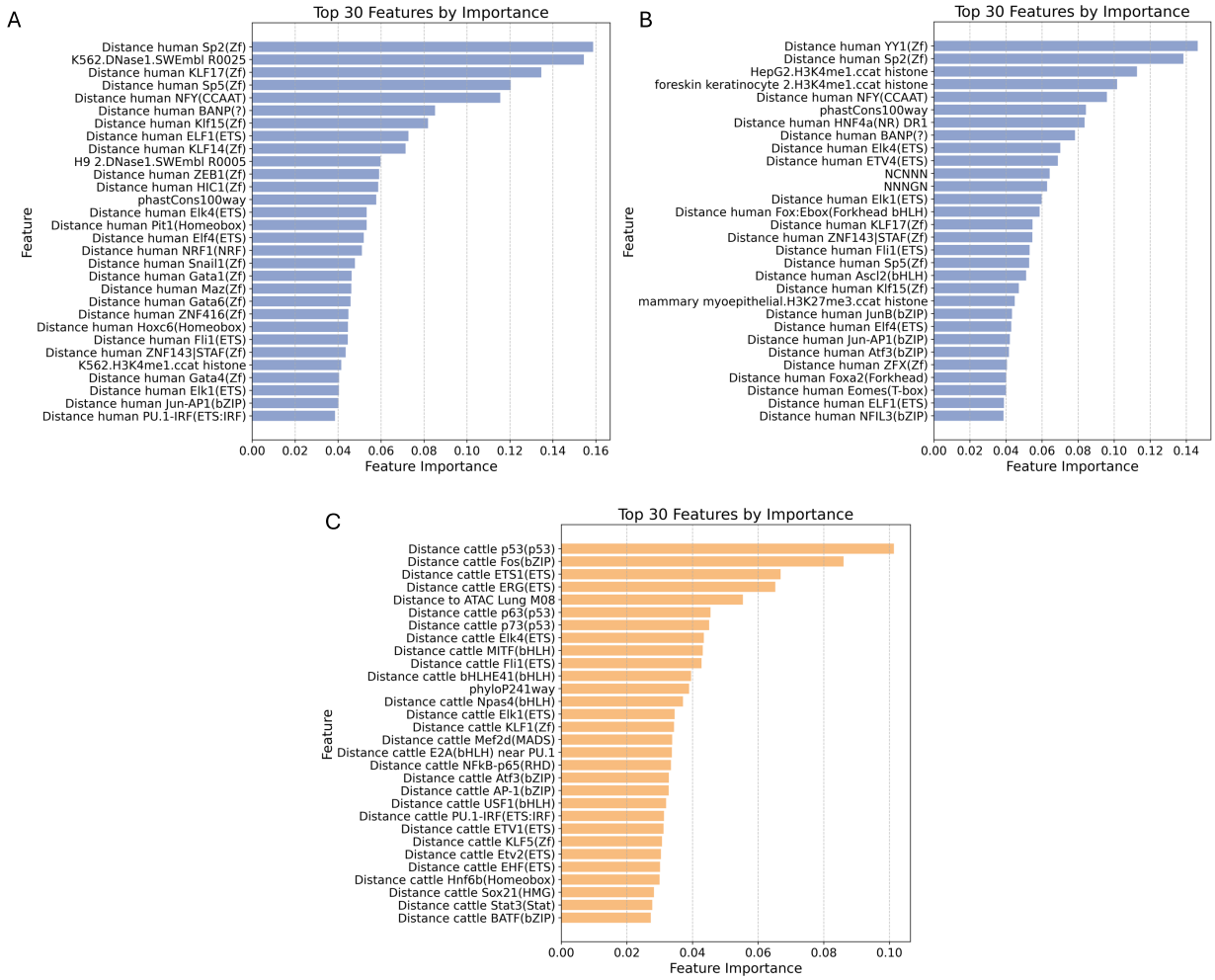
Supplementary Figure 9 ROC plots and AUROC scores for human models trained and tested with the same number of variants as the corresponding cattle models.



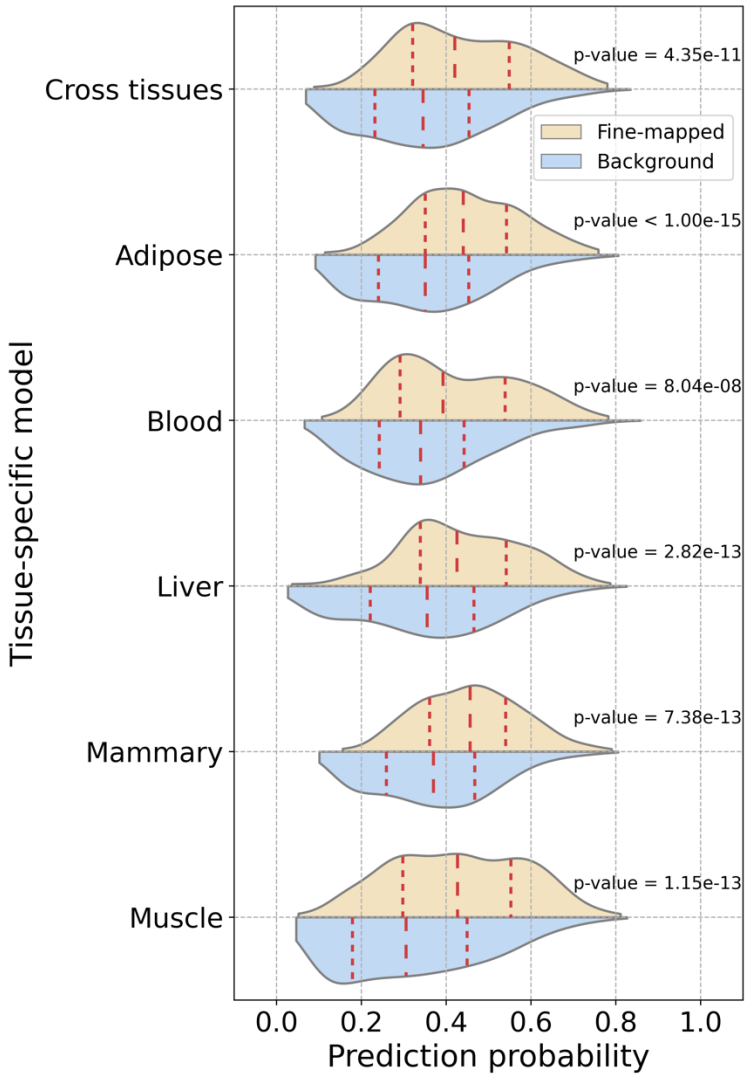
Supplementary Figure 10 The comparison of the AUROC scores of cattle tissue-specific models trained with or without the full set of Enformer features. The grey dashed line represents parity. The X and Y axis are limited from 0.60 to 0.80 for better clarity of the scatter plot.



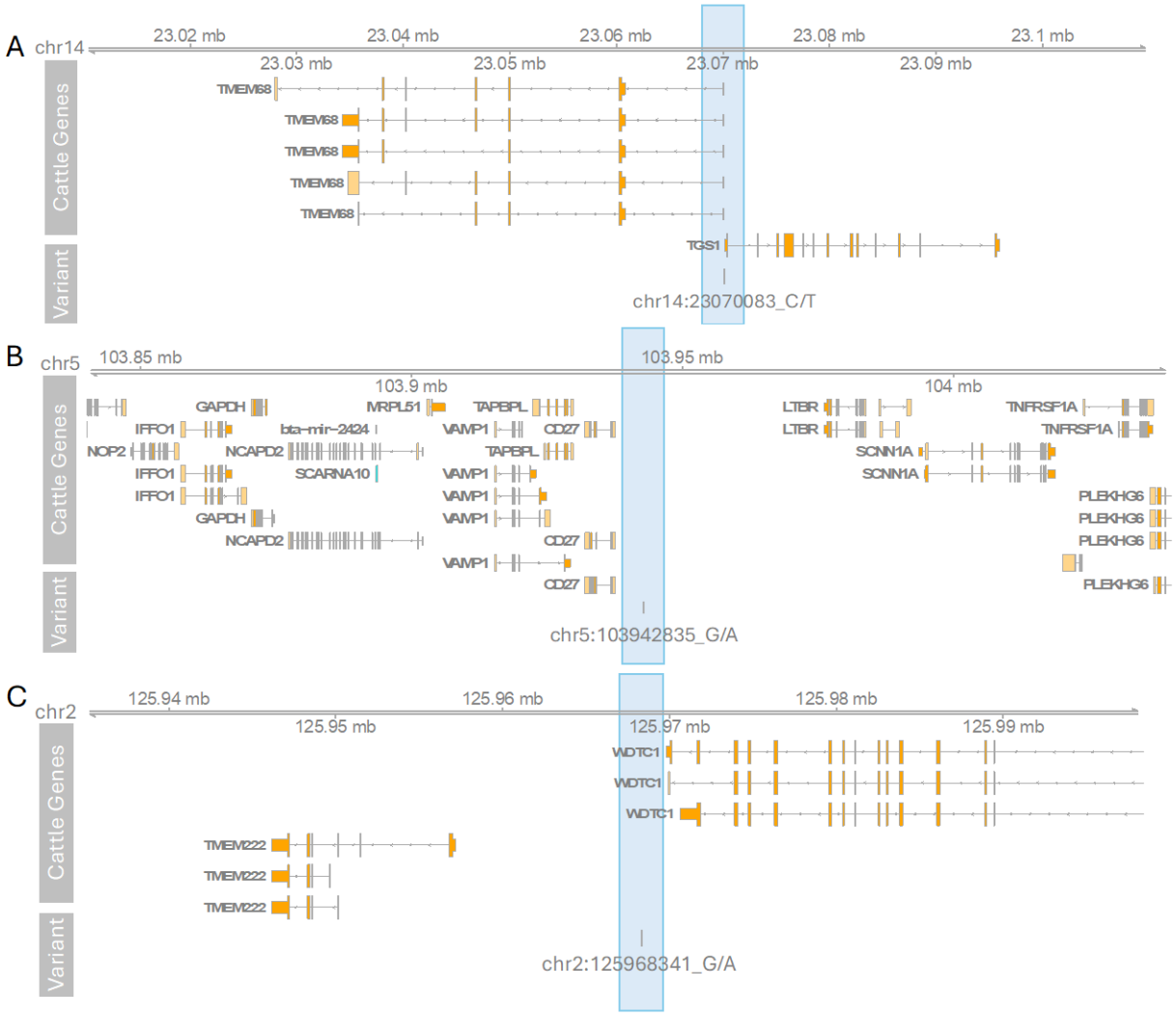
Supplementary Figure 11 ROC plots and AUROC (AUC) scores for human raQTL models trained and tested in (A) HepG2 cell line. (B) K562 cell line, and (C) both cell lines.



Supplementary Figure 12 Top 30 features in human and cattle raQTL models (A) K562 cell line, (B) HepG2 cell line and (C) cattle endothelial cell line. The features are ranked in descending order according to their relative feature importance in the model.



Supplementary Figure 13 Distributions of model derived prediction probabilities of cattle GWAS fine-mapped and background variants for stature. 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).



Supplementary Figure 14 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. (2021a). 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

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021b). 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

Benegas, G., Albors, C., Aw, A. J., Ye, C., & Song, Y. S. (2025). A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, 1–6. https://doi.org/10.1038/s41587-024-02511-w

Benegas, G., Eraslan, G., & Song, Y. S. (2025). *Benchmarking DNA Sequence Models for Causal Regulatory Variant Prediction in Human Genetics* (p. 2025.02.11.637758). bioRxiv. https://doi.org/10.1101/2025.02.11.637758

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

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLOS Computational Biology*, *6*(12), e1001025. https://doi.org/10.1371/journal.pcbi.1001025

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

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017b). 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

Duttke, S. H., Guzman, C., Chang, M., Delos Santos, N. P., McDonald, B. R., Xie, J., Carlin, A. F., Heinz, S., & Benner, C. (2024). Position-dependent function of human sequence-specific transcription factors. *Nature*, *631*(8022), 891–898. https://doi.org/10.1038/s41586-024-07662-z

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

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

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

Jiang, J., Cole, J. B., Freebern, E., Da, Y., VanRaden, P. M., & Ma, L. (2019). Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. *Communications Biology*, *2*(1), 1–12. https://doi.org/10.1038/s42003-019-0454-y

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

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

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

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

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. (2016a). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016b). 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

Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., Rowan, T. N., Low, W. Y., Zimin, A., Couldrey, C., Hall, R., Li, W., Rhie, A., Ghurye, J., McKay, S. D., Thibaud-Nissen, F., Hoffman, J., Murdoch, B. M., Snelling, W. M., ... Medrano, J. F. (2020). De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, *9*(3), giaa021. https://doi.org/10.1093/gigascience/giaa021

Santana, B. F., Riser, M., Hay, E. H. A., & Fragomeni, B. de O. (2023). Alternative SNP weighting for multi-step and single-step genomic BLUP in the presence of causative variants. *Journal of Animal Breeding and Genetics*, *140*(6), 679–694. https://doi.org/10.1111/jbg.12817

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

Schubach, M., Maass, T., Nazaretyan, L., Röner, S., & Kircher, M. (2024). CADD v1.7: Using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Research*, *52*(D1), D1143–D1154. https://doi.org/10.1093/nar/gkad989

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

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

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., Fang, L., Liu, S., Macleod, I. M., Liu, Z., Breen, E. J., Gao, Y., Liu, G. E., Tenesa, A., CattleGTEx Consortium, Mason, B. A., Chamberlain, A. J., Wray, N. R., & Goddard, M. E. (2023). Gene expression and RNA splicing explain large proportions of the heritability for complex traits in cattle. *Cell Genomics*, *3*(10), 100385. https://doi.org/10.1016/j.xgen.2023.100385

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