Journal Title Here, 2023, pp. 1–11

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Sequence-based chromatin activity modeling and regulatory impact prediction of genetic variants in farmed animals using deep learning

Dat Thanh Nguyen,[1,*, †] Tim Martin Knutsen,[2] Simen R. Sandve,[1] Sigbjørn Lien[1] and Lars Grønvold[1]

[1]Centre for Integrative Genetics, Faculty of Biosciences, Norwegian University of Life Sciences, 1432, Ås, Norway, [2]AquaGen AS, , P. O. Box 1240, NO-7462, Trondheim, Norway and [†]Current address: Centre for Precision Psychiatry, Institute of Clinical Medicine, University of Oslo, 0424, Oslo, Norway
[*]Corresponding author: Dat Thanh Nguyen, Centre for Integrative Genetics, Faculty of Biosciences, Norwegian University of Life Sciences Email:n.dat@outlook.com
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Non-coding genomic variations are crucial for the genetic regulation of traits; however, their functional impact in farmed animals remains underexplored due to limited genomic resources and the absence of tailored computational tools. Here, we present a deep learning-based framework that utilizes functional genomics data to generate genome-wide predictions of the regulatory impact of non-coding variants in cattle, chicken, pig, and Atlantic salmon. By leveraging chromatin profiles such as ATAC, DHS, and ChIP-seq data, we train and optimize separate deep networks for each species, achieving robust sequence modeling accuracy specific to each. Motif analysis confirms that the models capture regulatory grammar, while in silico saturation mutagenesis experiments provide meaningful interpretations of the functional impact of putative causal variants. Furthermore, functional scores derived from these models predict eQTL causal variants and enhance genomic prediction performance. Our findings highlight the transformative potential of sequence to function models in prioritizing causal variants and improving genomic prediction for livestock and aquaculture animals.

Key words: regulatory impact, deep learning, farmed animals, variant impact prediction, causal variant

Introduction

Non-coding genomic variations are known to constitute the majority of disease- and complex trait-associated single nucleotide polymorphisms (SNP) [1, 2, 3]. Despite the success in identifying these variants, it remains challenging to accurately determine which ones are causal based solely on association results, as many neutral genomic variants are also significantly associated with traits in GWAS due to linkage disequilibrium (LD) [4]. To complement the limitations of population-based association studies, high-throughput functional assays of regulatory elements such as DNase I hypersensitive sites (DHS) [5], transposase-accessible chromatin with sequencing (ATAC-seq) [6], and chromatin immunoprecipitation sequencing (ChIP-seq) [7] have become invaluable tools for detecting cis-regulatory elements (CRE) and prioritizing putative non-coding causal variants.

The advancement of deep learning, combined with the availability of large-scale functional annotations from projects like ENCODE [8] and Roadmap Epigenomics [9], has transformed the training of deep models for predicting chromatin states and assessing the impact of non-coding variants from DNA sequences in human genomics. In principle, deep neural networks are initially trained to differentiate putative regulatory sequences from background DNA sequences [10, 11, 12], or to directly predict experimental read coverage from high-throughput functional assays [13, 14]. Subsequently,

these models are applied to predict the regulatory impact of any genetic variants by analyzing the disparities between reference and alternative alleles. Due to rapid advancements in artificial intelligence, deep models for regulatory genomics have evolved into various architectures that continuously improve modeling performance. Notable architectures include DeepSEA and Basset, which are purely convolutional neural networks (CNNs) [10, 12]; Basenji, which employs dilated CNNs [13]; DanQ and DeepATT, which are hybrid CNNs with recurrent neural networks (RNNs) [11, 15]; and DeepFormer and Enformer, which are hybrid CNNs with transformers [14, 16].

In the realm of livestock and aquaculture genomics research, understanding the functional impact of non-coding variants is equally important. Variants in cis-regulatory elements can significantly influence transcriptional regulation, affecting traits of economic importance [17]. However, the exploration of functional non-coding variants in these species has fallen behind human genomics, primarily due to the absence of specialized functional genomics resources and tailored computational tools.

Genomic prediction has revolutionized livestock breeding by enabling the selection of individuals based on dense genomic marker information, thus increasing the precision of breeding value predictions for economically important traits [18, 19, 20, 21]. Despite its transformative success, genomic prediction primarily relies on SNPs as markers without considering the functional impact of the variants which potentially help to improve genomic prediction performance. The reason for this stem from the lack of effective methods for prioritizing functional variants, highlighting the urgent need for approaches to predict the impact of regulatory sequence variation in farmed animals.

In this work, we develop deep learning-based sequence models that utilize large-scale functional genomic datasets from multiple species, including cattle, pig, chickens, and Atlantic salmon. These datasets feature ATAC, DHS, histone modification, and transcription factor ChIP-seq profiles [22, 23] to prioritize functional variants. We evaluate variant effect predictions using expression quantitative trait loci (eQTL) data from large consortium projects, demonstrating that the models effectively predict putative functional variants. The predicted scores further enhances genomic prediction as illustrated in a case study with Atlantic salmon. Overall, our proposed functional scores show significant potential for prioritizing causal variants and enhancing genomic prediction practices in livestock and aquaculture breeding.

Methods

Datasets

An overview of the study is shown in Figure 1, which comprises three main steps: (i) data collection and preprocessing, (ii) deep learning model training and optimization, and (iii) regulatory impact inference. First, chromatin profile peaks from multiple farmed animal species are collected and preprocessed to obtain sequences and labels for deep model training (Figure 1 A, B). Next, we evaluate the performance of two widely used deep learning architectures for regulatory sequence modeling and variant impact prediction, DeepSEA [10] and DanQ [11], by testing different learning rates and applying the commonly used hold-out chromosome validation approach to identify the optimal model for each species (Figure 1 C). Finally, the best-performing models are used to infer the impact of regulatory variants for the corresponding species (Figure 1 D). Further details are provided in the following sections.

Training data

To train the machine learning models, we gather chromatin feature peaks from cattle (Bos taurus), pig (Sus scrofa), chicken (Gallus gallus), and Atlantic salmon (Salmo salar), sourced from the FAANG [22] and AQUA-FAANG [23] projects. For cattle, pig, and chicken, we utilize 95, 80, and 97 chromatin profiles, respectively, including data on histone modifications, CTCF ChIP-seq, DHS, and ATAC-seq across eight tissues: liver, lung, spleen, skeletal muscle, subcutaneous adipose, cerebellum, brain cortex, and hypothalamus. For Atlantic salmon, we use 301 profiles from histone modification, ChIP-seq, and ATAC-seq experiments across six tissues: brain, liver, gill, gonad, and muscle. Detailed statistics are provided in Figure 1 A.

For each species, we start by downloading the appropriate genome assembly version used for peak calling from the Ensembl database and then divide the genome into 200-bp bins starting from the first nucleotide of each chromosome. Specifically, we utilize the ARS-UCD1.2 genome for cattle, Sscrofa11.1 for pig, GalGal6 for chicken, and Ssal_v3.1 for Atlantic salmon. DNA sequences are extracted from each bin and encoded using a one-hot scheme: A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1], and N = [0,0,0,0]. To meet the input context length required by DeepSEA and DanQ, we include the sequences of two bins upstream and downstream of each target bin, concatenating them to create a final 1000-bp training sequence. The 200 bp bin for labeling ensures precise annotation of chromatin features, while the larger input window is designed to capture broader sequence dependencies and context, which is essential for learning complex regulatory grammar across diverse genomic backgrounds [10]. To label each DNA bin, we use BEDTools [24] to evaluate overlaps between bin coordinates and chromatin peak regions. Following the DeepSEA approach [10], a bin is assigned a label of 1 if at least 50% of its length overlaps with a peak region; otherwise, it is assigned 0. This process generates a label vector for each 200-bp bin, with the vector size corresponding to the number of chromatin profiles, as illustrated in Figure 1 B. Overall, the data processing pipeline generates approximately 17.6, 8.0, 17.3, and 15.7 million examples, with average positive label fractions of 1.51%, 2.96%, 2.67%, and 0.59% for cattle, chicken, pig, and salmon, respectively.

All autosomes, excluding those designated for validation and testing, are used for training. For cattle, chicken, and pig, held out validation and test sets are fixed, non-overlapping chromosome splits to ensure data independence and avoid leakage, similar to previous studies in model species [10, 12, 11]. Chromosome 21 is used for validation and chromosome 25 for testing in cattle and chicken, while chromosomes 16 and 17 are used for validation and testing, respectively, in pig. For salmon, due to the whole-genome duplication event resulting in highly similar genomic regions [25], random chromosome selection alone can not avoid data leakage during training. Instead, chromosomes 21 and 25, which are known to be duplicated, are kept for validation and testing.

Model architectures

DeepSEA [10] utilizes a three-block convolutional architecture for feature extraction. The first convolutional block consists of 320 filters with a kernel size of 8, followed by ReLU activation,

A. Datasets

Types	Cattle	Chicken	Pig	Salmon
ATAC/DNaseSeq	15	17	16	48
CTCF/DMC1	16	16	0	56
H3K27ac	16	16	16	47
H3K27me3	16	16	16	47
H3K36me3	0	0	0	9
H3K4me1	16	16	16	47
H3K4me3	16	16	16	47
Sum	**95**	**97**	**80**	**301**

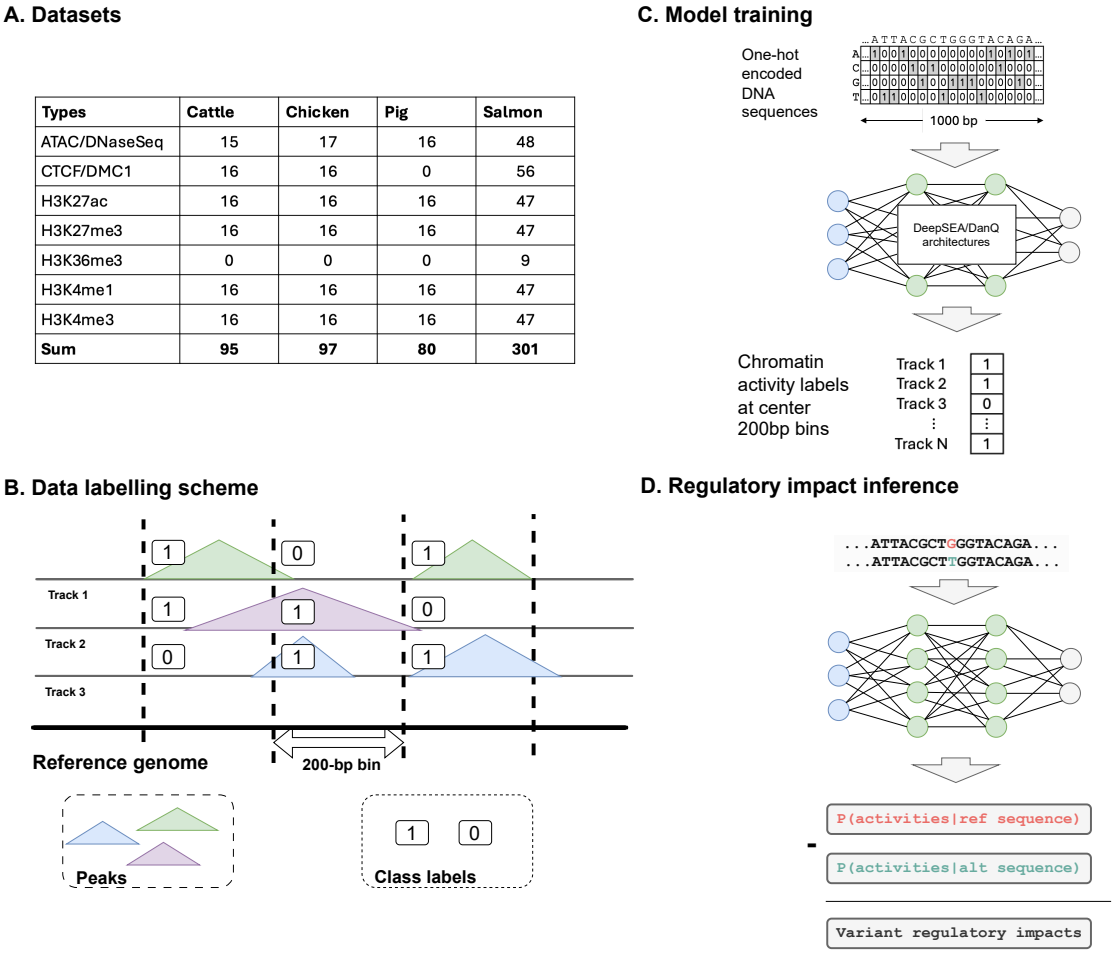

Fig. 1. Overview of the study. (A) Summary of chromatin profiles used for each species. (B) Data preprocessing workflow for generating sequence-level labels. (C) Deep learning model training and optimization, using a one-hot encoded DNA sequence matrix of size 1000×4 as input and producing a binary label vector of length equal to the number of chromatin profiles as output.

(D) Variant impact inference scheme.

max pooling with a window size of 4, and dropout with a probability of 0.2. The second block extends this architecture with 480 filters of the same size, applying identical pooling operations and maintaining the dropout rate of 0.2. The third block further increases the complexity by incorporating 960 filters, and includes a higher dropout rate of 0.5 to better address over fitting. Following the convolutional and pooling operations, the network flattens the output and processes it through two fully connected layers to produce the final classification outputs.

DanQ [11] begins with a convolutional block that applies 320 filters of size 26 to the input sequences, followed by ReLU activation, max pooling with a window size of 13, and dropout with a probability of 0.2. The output of this block is then fed into a bidirectional LSTM (Long Short-Term Memory) layer, which has 320 hidden units and processes the sequence data to capture temporal dependencies. Dropout with a probability of 0.5 is applied to the LSTM output to prevent over fitting. The output is then flattened and passed through two fully connected layers for classification.

Model training and evaluation

Deep learning models are trained using TensorFlow v2.12.0 on an NVIDIA Quadro RTX 8000 GPU, with parameters randomly initialized using TensorFlow's default settings. For optimization, we apply the widely used procedure for training and evaluating deep models for epigenetic sequence modeling [10, 26, 27]. Binary Cross-Entropy (BCE) loss function is used as the objective function defined as $\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$, where y_i is the true label, $\hat{y}_i$ is the predicted probability, and N is the number of samples. The Adam optimizer [28] is applied with a weight decay of 1×10^{-6} and a batch size of 1024. To identify the optimal learning rate for each method-dataset pair, we evaluate four candidate values: 1×10^{-3}, 5×10^{-4}, 1×10^{-4}, and 5×10^{-5}. Models are trained for up to 100 epochs, with early stopping applied if the validation loss does not improve for 5 consecutive epochs.

Finally, model performance is assessed on hold-out test sets. Both the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPR) are used to select the best model for each species.

These metrics are particularly appropriate for imbalanced classification problems, where AUROC evaluates the trade-off between true positive rate (TPR) and false positive rate (FPR), while AUPR focuses on the model's ability to correctly identify positive instances, emphasizing precision and recall [29]. Both forward and reverse DNA sequences are used during training, while only forward sequences are considered for validation and final evaluation.

Regulatory impact prediction and in silico saturated mutagenesis

Regulatory impact prediction involves evaluating the potential effects of genetic variants on regulatory activities [30, 10, 12]. For each variant, two 1000-bp sequences centered on the variant position are extracted from the reference genome: one carrying the reference allele and the other carrying the alternative allele. These sequences are used as input to the trained model, which calculates the probabilities $P_{\text{reference}}$ and $P_{\text{alternative}}$ that the reference and alternative sequences belong to active chromatin regions, respectively. The regulatory impact of the variant is inferred as the difference between these probabilities: $\Delta P = |P_{\text{reference}} - P_{\text{alternative}}|$.

In silico saturation mutagenesis extends regulatory impact prediction by evaluating the effect of all possible single-nucleotide substitutions within a predefined genomic region, rather than being limited to known variants. In details, for each position in the region, all possible alternative nucleotides are substituted into the reference sequence, generating a set of mutated sequences. Each sequence is then passed through the trained model to obtain predicted probabilities of regulatory activity. The effect of a mutation is quantified as the absolute difference in predicted activity between the mutated and reference sequences. This procedure yields a high-resolution map of nucleotide-level regulatory sensitivity, enabling identification of functionally critical sites across the region [12].

Motif Analysis

We perform motif analysis by extracting the convolutional filters from the first layer of each trained model and converting them into position probability matrices (PPMs). This is achieved by applying a softmax normalization across the four nucleotides at each position, which converts filter weights into probabilities while preserving relative weight contributions and ensuring each position's probabilities sum to 1.

For a filter weight matrix $W \in \mathbb{R}^{4\times m}$ (where 4 represents nucleotides and m is filter length), the conversion for nucleotide i at position j is given by: $\text{PPM}_{ij} = \frac{\exp(W_{ij})}{\sum_{k=1}^{4}\exp(W_{kj})}$.

We then compare the resulting PPMs against known transcription factor binding motifs using the TOMTOM algorithm [31], implemented through the MEME Suite web server [32]. We use the JASPAR 2022 CORE non-redundant vertebrates v2 database [33] as the reference motif set.

Fine-mapped eQTL variant classification

To assess the utility of the predicted functional scores, we adopt the methodology described by Avsec et al. [14], focusing on putative eQTL causal variants. Specifically, we utilize the SuSiE [34] fine-mapped results across multiple tissues provided by the PigGTEx consortium [35]. We include only tissues with at least 500 variants having a posterior inclusion probability (PIP) exceeding 0.9 in a credible causal set, resulting in 13 tissue-specific causal variant datasets. Negative sets are matched by sampling variants with PIP < 0.01 but $|Z\text{-score}| > 4$ for the same gene. If no such variants are available, alternatives are selected from genome-wide variants with PIP < 0.01 and $|Z\text{-score}| > 6$.

Using predicted functional scores, we then train separate random forest classifiers for each tissue to distinguish between positive and negative variant sets using ten-fold cross-validation. Default hyperparameters from scikit-learn are used, and 50 iterations of stochastic cross-validation and random forest fitting are performed to estimate model accuracy and its standard deviation.

Genomic selection analysis

As genomic prediction is a core application of livestock and aquaculture genomics, we further evaluate the potential of functional scores for prioritizing SNP sets within genotyping arrays. The Atlantic salmon SNP array, developed by AquaGen, comprises 70,000 markers (70k array) is downscaled to 9,073 SNPs, corresponding to a density of one SNP per 250 kb. Seven SNP sets are constructed: five randomly selected subsets from the 70k array, one functional set comprising the top 9,073 SNPs with the highest functional scores across the array, and one functional set consisting of the top 9,073 SNPs with the highest functional scores stratified by 250 kb windows (i.e., selecting the SNP with the highest functional score within each bin). Functional scores for ranking SNPs are computed as the average scores derived from epigenomic profiles of brain, gonad, and liver tissues of male Atlantic salmon.

Two widely utilized genomic prediction methods are applied including GCTA-Yang, a SNP-BLUP approach implemented in GCTA; and GCTB-Bayesian, a Bayesian framework implemented in GCTB. The GCTA-Yang method adopts a genomic best linear unbiased prediction (GBLUP) framework, assuming equal contributions of all SNPs to the genetic architecture of the trait and leveraging a genomic relationship matrix to predict genetic values [36]. Conversely, the GCTB-Bayesian method fits all genotyped markers simultaneously, accommodating heterogeneity in trait genetic architecture and providing a more flexible framework for genome-wide association studies and genomic prediction [37].

The dataset used in this analysis comprises Atlantic salmon from three cohorts: AS19, AS20, and AS21. Each individual is genotyped for SNPs and assigned a binary phenotype representing late maturation. Covariates, such as year and age are included to account for environmental and batch effects. The models are trained using data from the combined AS19 and AS20 cohorts and validated on the genetically distinct AS21 cohort, ensuring a robust assessment of model generalizability.

Training is conducted across all SNP sets and prediction methods. Correlations between observed phenotypes and predicted phenotypes are subsequently calculated for each SNP set, providing a quantitative measure of the effectiveness of functional scores in enhancing genomic prediction.

Results

Optimizing chromatin activity modeling for farmed species

To optimize chromatin activity modeling, we evaluate two widely used deep learning architectures, DeepSEA and DanQ, across multiple learning rates. The results are shown in Figure 2 and S.1, and Table S.1 and S.2. Overall, both architectures demonstrate high performance across species,

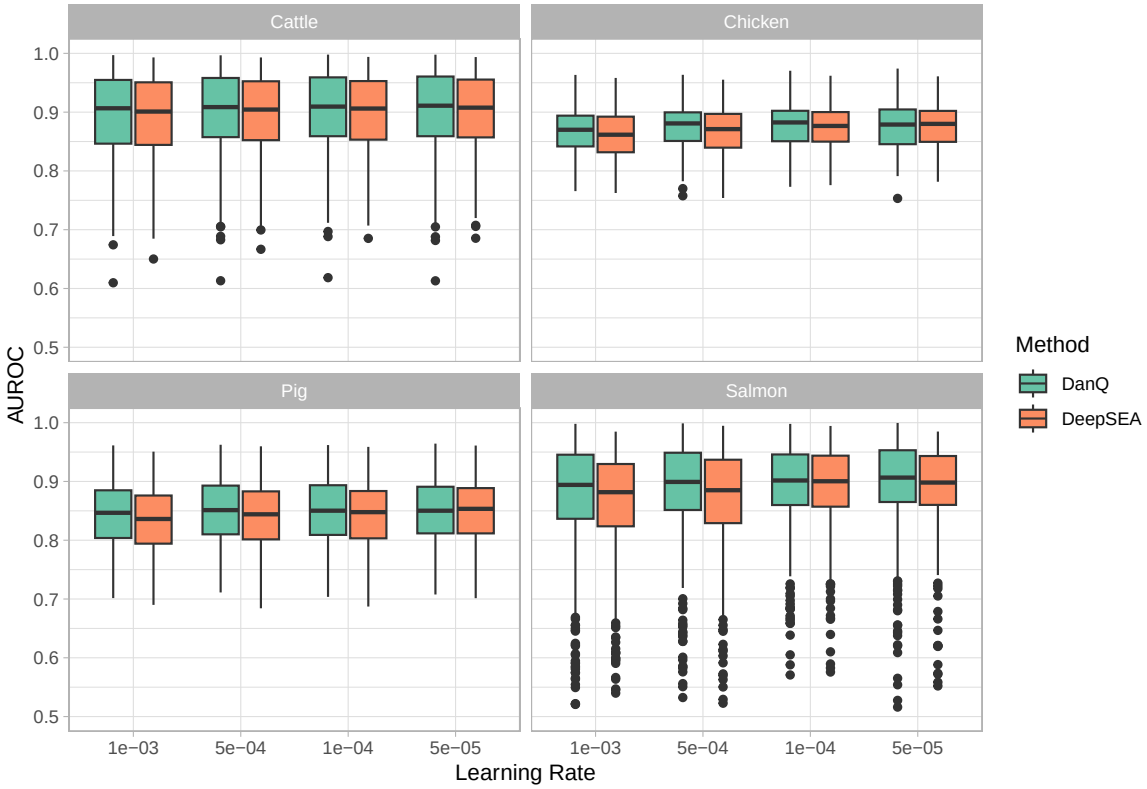

Fig. 2. Performance comparison (AUROC scores) across learning rates of DeepSEA and DanQ in four species: cattle, chicken, pig, and salmon.

achieving robust AUROC scores. Notably, DanQ dominantly outperforms DeepSEA with slightly higher AUROC scores across species and learning rates tested. For instance, in Cattle, the median AUROC score for DanQ ranges from 0.9066 to 0.9110, while DeepSEA achieves scores between 0.9010 and 0.9076. Similarly, in chicken, DanQ scores range from 0.8700 to 0.8825, compared to 0.8616 to 0.8800 for DeepSEA. The advantage of DanQ may be attributed to its ability to model the temporal dynamics of regulatory vocabularies [11]. The AUPR metric further supports this observation. In cattle, DanQ achieves median AUPR scores ranging from 0.4389 to 0.4574, compared to 0.4305 to 0.4561 for DeepSEA. In chicken, DanQ scores range from 0.4544 to 0.4859, while DeepSEA ranges from 0.4171 to 0.4670. For pig, DanQ achieves 0.3255 to 0.3400, slightly higher than DeepSEA's 0.3107 to 0.3333. In salmon, where predictive performance is generally lower, DanQ achieves AUPR scores ranging from 0.0993 to 0.1288.

Building upon these findings, we choose DanQ as the primary method for evaluation by selecting the best-performing DanQ model for each species (Table S.3). For cattle, the DanQ model trained with a learning rate of 5×10^{-5} achieves a median AUROC score of 0.9110 and a corresponding median AUPR of 0.4574. Similarly, in chicken, the best DanQ model, trained with a learning rate of 1×10^{-4}, records a median AUROC score of 0.8825 and a median AUPR of 0.4846. For pig, the highest-performing DanQ model utilizes a learning rate of 5×10^{-4}, achieving a median AUROC of 0.8512 with a median AUPR of 0.3362. Lastly, the salmon DanQ model trained at a learning rate of 5×10^{-5} achieves a median AUROC score of 0.9065 and a corresponding median AUPR of 0.1288.

To gain deeper insights, we explore the variability in performance across epigenetic profiles. As shown in Figure S.2, the AUROC scores exhibit significant variation depending on the data type and species. For instance, in cattle, H3K27ac and CTCF data achieve the highest median AUROC scores, while in salmon, H3K4me3 and ATAC/DNase stand out as the best-performing features. Chicken and pig display intermediate performance levels across most features. These results suggest that certain epigenetic features are more informative than others for regulatory activity modeling in the different species. Overall, these results indicate that we have achieved high accuracy, but the effectiveness of regulatory activity modeling varies across species and type of epigenetic features.

Deep models learn transcription factor binding motifs

To verify that the deep models effectively learn regulatory grammar, we conduct motif analysis. Following the approach outlined in the method section, for each species, we convert the weight matrices from the first convolutional layer of the trained deep models into PPMs. These learned motifs are then compared against the JASPAR 2022 CORE non-redundant v2 database using the TOMTOM algorithm [31].

Among the 320 motifs learned by each DanQ model, 49, 51, 63, and 34 motifs significantly match known motifs in the target database (E-value < 0.01) for cattle, chicken, pig, and salmon, respectively. Among the significant matched motifs in cattle, we select four representative motifs and visualize them for illustration, as shown in Figure 3. These findings suggest that the trained models effectively learn regulatory vocabulary. However, the number of matched motifs is substantially lower than the 166 out of 320 matches reported for the human DanQ

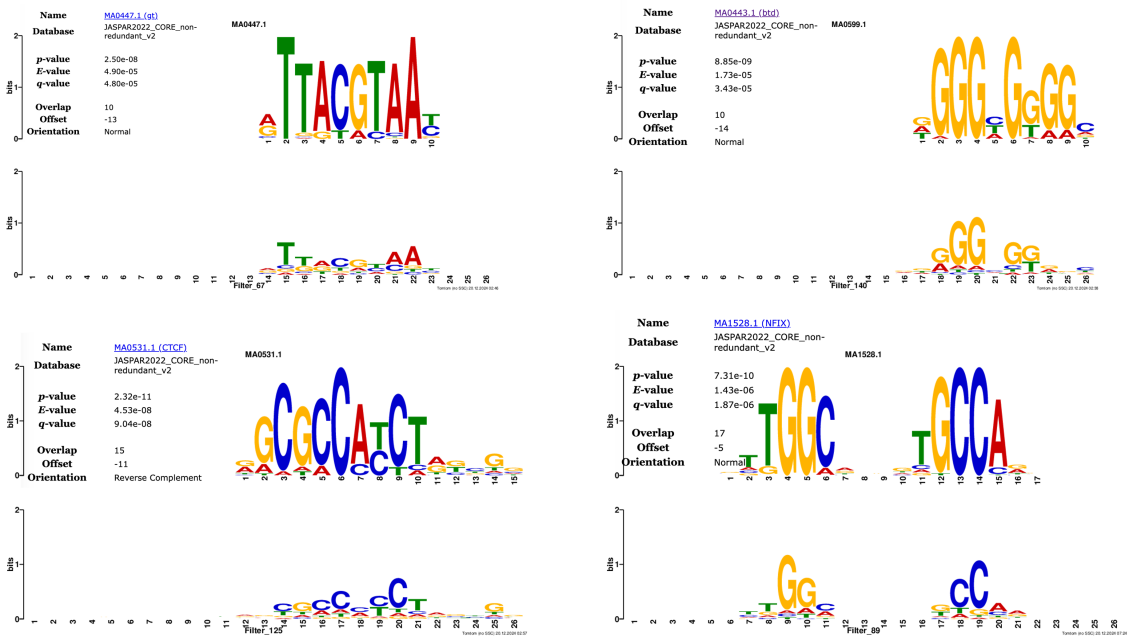

Fig. 3. Visualization of four motifs GT, BTD, CTCF, and NFIX in cattle.

model [11]. This disparity may reflect the under-representation of farmed animal motifs in current databases and the limitation in term of training data for these species.

Functional scores predict eQTL causal variants

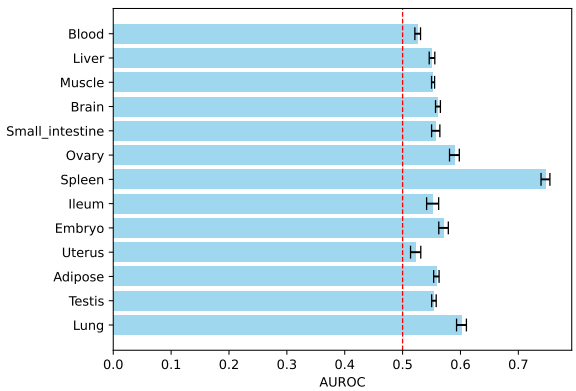

Fig. 4. Mean and standard deviation of AUROC scores for 50 random forest eQTL causal variant classifiers across 13 pig eQTL datasets.

The utility of the predicted functional scores is evaluated through random forest classification across 13 pig tissue-specific datasets, distinguishing putative eQTL causal variants from matched negative variants. The mean AUROC scores from 50 iterations of ten-fold cross-validation vary across tissues, highlighting differences in predictive performance. The AUROC scores range from 0.7469 in spleen, which demonstrates the strongest predictive ability, to 0.5226 in uterus, indicating lowest performance. Other tissues exhibit intermediate levels of predictive accuracy, with scores of 0.6018 (lung), 0.5541 (testis), 0.5583 (adipose), 0.5707 (embryo), 0.5519 (ileum), 0.5895 (ovary), 0.5572 (small intestine), 0.5611 (brain), 0.5526 (muscle), 0.5508 (liver), and 0.5261 (blood).

The overall mean AUROC score across all tissues is 0.5726, indicating moderate performance of the functional scores in predicting putative eQTL causal variants. The results, including the mean and standard deviation of model performance, are detailed in Figure 4. These findings suggest that the predicted functional scores are correlated with eQTL predicted causal variants, although the strength of correlation varies across tissues.

In silico saturation mutagenesis aids functional interpretation of putative causal variants

A trained model can be used to predict the functional activity of any given sequence, providing a powerful method for understanding and utilizing the regulatory patterns it has learned. In silico saturation mutagenesis experiments, which involve testing every possible mutation in a sequence, serve as an effective tool for identifying the specific nucleotides responsible for functional activity [12]. This approach is similar to standard functional impact prediction, where the difference in probabilities between two genotypes is computed. However, the inference is not limited to the variant position but extends to all possible variants within the targeted region.

We apply this strategy to investigate the putative regulatory variant rs133257289 in cattle, which is the top colocalized SNP of the cis-eQTL of DGAT1 and the GWAS association with protein yield [38, 39]. In Figure 5, we present heat maps showing the change in predicted chromatin accessibility (mean of liver ATAC-seq tracks) due to mutations at each position, replacing the original nucleotide with each alternative for sequences surrounding the variant. These maps highlight the nucleotides most crucial to a sequence's activity. For each position, we assign two scores: the loss score, which measures the largest possible decrease in activity, and the gain score, which measures the largest increase.

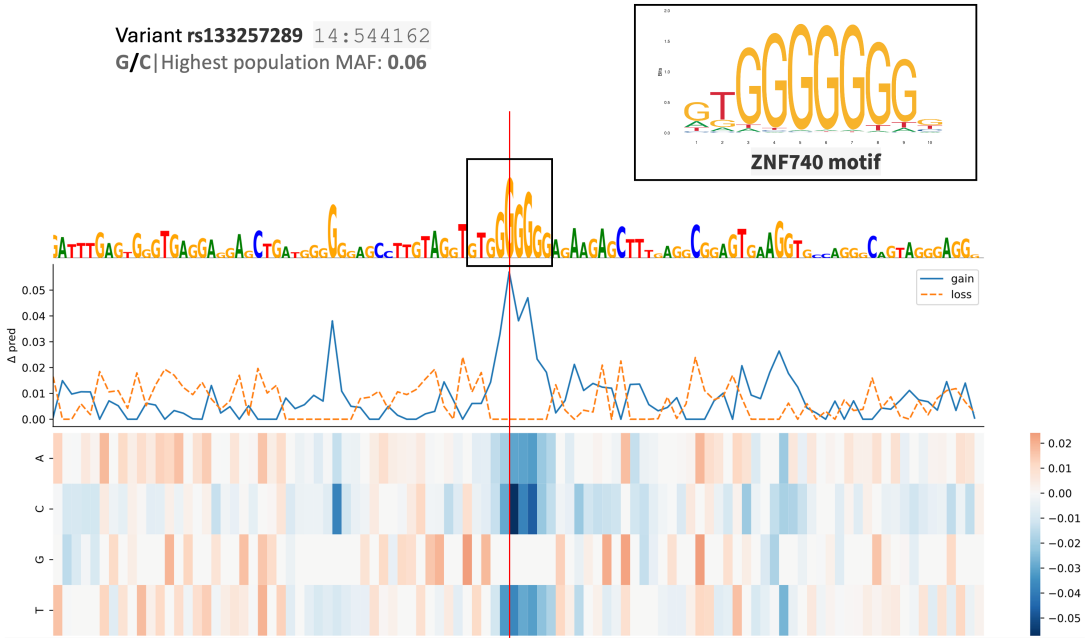

Fig. 5. In silico saturation mutagenesis experiments of rs133257289, a putative regulatory variant in cattle, which is the top colocalized SNP for the cis-eQTL of DGAT1 and the GWAS association with protein yield.

High gain scores correspond to positions associated with the ZNF740 motif, where mutations disrupt the motif and increase chromatin accessibility. Notably, a G to C mutation aligns with the observed effect size of increased gene expression in the corresponding eQTL test in the liver (p-value $< 1.2 \times 10^{-21}$, effect size = 0.3123) [38]. These findings highlight the potential of in silico saturation mutagenesis experiments using trained models for interpreting putative causal variants.

Functional impact scores improve genomic selection

To assess the impact of SNP selection strategies on genomic prediction accuracy, we evaluate the predictive correlation between observed and predicted phenotypes across various SNP sets and genomic prediction methods (Figure 6 and Table S.4). The SNP sets include the full 70k SNP array, the top 9,073 SNPs selected based on functional scores (both genome-wide and within 250 kb bins), and random subsets of 9,073 SNPs derived from the 70k array.

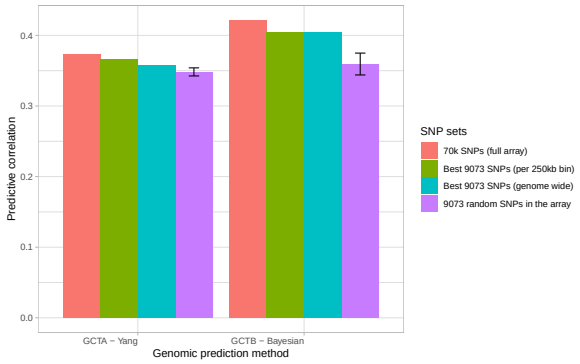

Fig. 6. Genomic prediction performance of various SNP sets.

For the GCTA-Yang method, which assumes equal SNP contributions, the full 70k SNP array exhibits the highest predictive performance (correlation = 0.373), followed closely by the functional SNP set stratified by 250 kb bins (correlation = 0.365). SNP sets based on genome-wide functional scores (correlation 0.358) and random subsets of the array (correlation = 0.348 ± 0.0057) show slightly lower predictive correlations. In contrast, the GCTB-Bayesian method demonstrates improved predictive accuracy compared to GCTA-Yang across all SNP sets. The 70k SNP array achieves the highest predictive correlation (mean = 0.421), with functional SNP sets (both genome-wide and per 250 kb bin) showing comparable predictive performance (correlation = 0.404 for both SNP sets). Random SNP subsets yield the lowest predictive correlation under this method (correlation = 0.359 ± 0.0155).

Overall, these findings highlight the potential utility of functional scores in prioritizing SNPs for genomic prediction. Specifically, SNP sets derived from functional annotations perform on par with the full array, while significantly outperforming random SNP subsets. The results also demonstrate the advantages of the GCTB-Bayesian method over the GCTA-Yang approach, particularly in scenarios involving the maturation trait in Atlantic salmon.

Discussion

Despite its promise, the application of deep learning for predicting the impact of genetic variants in farm animals remains underexplored [40, 41]. This study presents four deep learning sequence-to-function models, each trained on diverse functional genomics data from cattle, pigs, chicken or salmon. We demonstrate the model's ability to learn regulatory motifs and predict regulatory impact of non-coding variants. By enabling more precise identification of functional variants, these models can support more informed breeding

decisions, ultimately contributing to sustainable improvements in livestock productivity and resilience. This work sets a foundation for further research and application of deep learning in the field of animal genomics.

We tried both the DeepSEA and DanQ deep learning architectures, settling on the DanQ architecture as it consistently returned higher AUROC scores in all species. This difference in performance between the architectures was also shown in the original DanQ model that was trained on human data [11], indicating the finding generalizes well across species. The key difference between these architectures is the recurrent bLSTM block used in DanQ which has the ability to learn positional dependencies in the sequence, possibly reflecting the nature of the regulatory grammar.

Our motif analysis highlights the biological relevance of these models, demonstrating their ability to learn sequence patterns corresponding to known transcription factor binding sites. The DanQ model, with its single convolutional layer and long filters, is particularly suited for such analyses, as these filters effectively function as motif detectors. We identified motifs various significant motifs in all species, underscoring the models' capability to capture regulatory elements even with limited training datasets in farmed animals. However, the reduced number of detected motifs compared to the human DanQ model may stem from key differences in the datasets. The original DanQ model was trained on a comprehensive functional data with large amount of TF ChIP-seq data making it to learn binding motifs specific to transcription factors in the dataset. In contrast, our models rely on broader chromatin accessibility profiles, which may limit motif specificity. Another factor to consider is the evolutionary divergence between species transcription factors in distant species compared to the reference database, such as salmon, may recognize motifs distinct from those in human or mouse, further influencing motif detection. These findings highlight the need for more comprehensive functional annotations in non-human species, which could enhance both motif discovery and overall model performance.

The application of in silico saturation mutagenesis demonstrates the practical utility of these models in interpreting putative causal regulatory variants. For instance, the analysis of rs133257289 in cattle reveals specific nucleotide changes that influence chromatin accessibility and align with observed eQTL effects, underscoring the potential of deep learning-based approaches for functional variant interpretation. Such methods provide invaluable insights for identifying candidate variants linked to economically important traits, paving the way for more targeted breeding strategies.

Moreover, our evaluation of functional scores in predicting eQTL causal variants in pig reveals moderate to high accuracy, with performance varying across tissues. Closer examination of the training data reveals that predictive performance is generally higher in tissues represented in the training chromatin profiles. Notably, spleen (0.7469), lung (0.6018), and liver (0.5508) which are included in the training data—exhibit relatively stronger performance compared to tissues such as uterus (0.5226), testis (0.5541), and blood (0.5261), which are absent from the training set. This suggests that the presence of tissue-matched chromatin features, including histone modifications, DHS, ATAC-seq, and CTCF binding profiles, enhances the classification of putative eQTLs. Intermediate AUROC values in some tissues may reflect partial similarity or shared regulatory architecture with training tissues. These findings highlight the importance of incorporating diverse, tissue-relevant epigenomic data to improve the generalizability and predictive power of functional variant scoring models.

Importantly, these functional scores demonstrate significant potential for enhancing genomic selection, as evidenced by their comparable performance to full SNP arrays and their superiority over randomly selected subsets as demonstrated in salmon. This finding suggests that functional annotations can be used to prioritize SNPs for functional-aware array design [42], thereby improving the efficiency and accuracy of genomic prediction.

Finally, we note that class imbalance is a common phenomenon in regulatory sequence modeling, where positive regulatory elements are typically outnumbered by negative examples. Our study adheres to standard practices widely adopted in the field, including those employed by models such as DeepSEA [10], DanQ [11], and PlantDeepSEA [27], which train the models with standard binary cross-entropy objective on genome-wide data. Nevertheless, a systematic investigation into the impact of class imbalance on model performance and solutions for improving generalization are still lacking and would be a valuable direction for future research.

Conclusion

Overall, this study demonstrates the feasibility and utility of deep learning-based approaches for predicting the regulatory impact of non-coding variants in farmed species. By leveraging functional genomic data, our models achieve high predictive accuracy, uncover regulatory grammar, and provide actionable insights for genomic selection. The integration of functional annotations into breeding programs offers a promising approach for enhancing the precision and efficiency of livestock and aquaculture genomics. Future work should focus on expanding functional genomics resources for farmed species and exploring the application of these methods to other economically important organisms.

Competing interests

TMK is employed by Aquagen AS. The remaining authors declare no conflict of interest.

Availability of data and materials

The deep learning models and full Nextflow codes to reproduce data preprocessing and training are available at `https://doi.org/10.6084/m9.figshare.28547183.v1` or `https://github.com/datngu/DeepFARM`. Trained model weights, data supporting the findings and source codes for data analyses and generating figures for this study are available at `https://github.com/datngu/DeepFARM_paper` or `https://doi.org/10.6084/m9.figshare.29433326.v1`. Functional datasets are available in its original publications [22, 23] at `http://farm.cse.ucdavis.edu`, and `https://salmobase.org/`. Cattle GTEX and Pig GTEX data are available at `https://cgtex.roslin.ed.ac.uk/`, and `https://piggtex.ipiginc.com/`. The datasets used in genomic prediction analysis are property of Aquagen AS that need permission to access.

Author contributions

DTN: conceptualized and implemented the methods; designed and analyzed the data; and wrote the manuscript with inputs from other co-authors. TMK: conceptualized and implemented the genomic prediction experiment. LG, SRS and SL: contributed to discussions, manuscript review, and project administration. All authors read and approved the manuscript.

Use of AI Software

Large language models were used to improve the wording and grammar of some texts, but not to generate new content.

Acknowledgments

The authors would like to thank the AQUA-FAANG consortium for granting access to the dataset, the Orion HPC for providing computational resources. DTN also gratefully acknowledges the financial support from internal funding scheme at Norwegian University of Life Sciences (project number 1211130114), which financed the international stay at Queensland Institute of Medical Research, Brisbane, Australia.

Funding

This work is supported by the NMBU Doctoral Research Fellowship.

References

1. Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.
2. Zhi-Liang Hu, Carissa A Park, and James M Reecy. Building a livestock genetic and genomic information knowledgebase through integrative developments of animal qtldb and corrdb. *Nucleic acids research*, 47(D1):D701–D710, 2019.
3. Dongmei Tian, Pei Wang, Bixia Tang, Xufei Teng, Cuiping Li, Xiaonan Liu, Dong Zou, Shuhui Song, and Zhang Zhang. Gwas atlas: a curated resource of genome-wide variant-trait associations in plants and animals. *Nucleic Acids Research*, 48(D1):D927–D932, 2020.
4. Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019.
5. Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.
6. Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218, 2013.
7. David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.
8. ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.
9. Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.
10. Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.
11. Daniel Quang and Xiaohui Xie. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, 44(11):e107–e107, 2016.
12. David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–999, 2016.
13. David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018.
14. Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.
15. Jiawei Li, Yuqian Pu, Jijun Tang, Quan Zou, and Fei Guo. Deepatt: a hybrid category attention neural network for identifying functional effects of dna sequences. *Briefings in bioinformatics*, 22(3):bbaa159, 2021.
16. Zhou Yao, Wenjing Zhang, Peng Song, Yuxue Hu, and Jianxiao Liu. Deepformer: a hybrid network based on convolutional neural network and flow-attention mechanism for identifying the function of dna sequences. *Briefings in Bioinformatics*, 24(2):bbad095, 2023.
17. Ross D Houston, Tim P Bean, Daniel J Macqueen, Manu Kumar Gundappa, Ye Hwa Jin, Tom L Jenkins, Sarah Louise C Selly, Samuel AM Martin, Jamie R Stevens, Eduarda M Santos, et al. Harnessing genomics to fast-track genetic improvement in aquaculture. *Nature Reviews Genetics*, 21(7):389–409, 2020.
18. Theo HE Meuwissen, Ben J Hayes, and ME1461589 Goddard. Prediction of total genetic value using genome-wide dense marker maps. *genetics*, 157(4):1819–1829, 2001.
19. Mario PL Calus. Genomic breeding value prediction: methods and procedures. *animal*, 4(2):157–164, 2010.
20. Luiz F Brito, Shannon M Clarke, John C McEwan, Stephen P Miller, Natalie K Pickering, Wendy E Bain, Ken G Dodds, Mehdi Sargolzaei, and Flávio S Schenkel. Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a hd snp chip. *BMC genetics*, 18:1–17, 2017.
21. Daniela AL Lourenco, Breno O Fragomeni, Shogo Tsuruta, Ignacio Aguilar, Birgit Zumbach, Rachel J Hawken, Andres Legarra, and Ignacy Misztal. Accuracy of estimated breeding values with genomic information on

males, females, or both: an example on broiler chicken. *Genetics Selection Evolution*, 47:1–9, 2015.

22. Colin Kern, Ying Wang, Xiaoqin Xu, Zhangyuan Pan, Michelle Halstead, Ganrea Chanthavixay, Perot Saelao, Susan Waters, Ruidong Xiang, Amanda Chamberlain, et al. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nature communications*, 12(1):1821, 2021.
23. Ian A Johnston, Matthew P Kent, Pierre Boudinot, Mark Looseley, Luca Bargelloni, Sara Faggion, Gabriela A Merino, Garth R Ilsley, Julien Bobe, Costas S Tsigenopoulos, et al. Advancing fish breeding in aquaculture through genome functional annotation. *Aquaculture*, page 740589, 2024.
24. Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.
25. Sigbjørn Lien, Ben F Koop, Simen R Sandve, Jason R Miller, Matthew P Kent, Torfinn Nome, Torgeir R Hvidsten, Jong S Leong, David R Minkley, Aleksey Zimin, et al. The atlantic salmon genome provides insights into rediploidization. *Nature*, 533(7602):200–205, 2016.
26. Kathleen M Chen, Evan M Cofer, Jian Zhou, and Olga G Troyanskaya. Selene: a pytorch-based deep learning library for sequence data. *Nature methods*, 16(4):315–318, 2019.
27. Hu Zhao, Zhuo Tu, Yinmeng Liu, Zhanxiang Zong, Jiacheng Li, Hao Liu, Feng Xiong, Jinling Zhan, Xuehai Hu, and Weibo Xie. Plantdeepsea, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. *Nucleic Acids Research*, 49(W1):W523–W529, 2021.
28. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
29. Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.
30. Dongwon Lee, David U Gorkin, Maggie Baker, Benjamin J Strober, Alessandro L Asoni, Andrew S McCallion, and Michael A Beer. A method to predict the impact of regulatory variants from dna sequence. *Nature genetics*, 47(8):955–961, 2015.
31. Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome biology*, 8:1–9, 2007.
32. Timothy L Bailey, James Johnson, Charles E Grant, and William S Noble. The meme suite. *Nucleic acids research*, 43(W1):W39–W49, 2015.
33. Jaime A Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, et al. Jaspar 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 50(D1):D165–D173, 2022.
34. Ran Cui, Roy A Elzur, Masahiro Kanai, Jacob C Ulirsch, Omer Weissbrod, Mark J Daly, Benjamin M Neale, Zhou Fan, and Hilary K Finucane. Improving fine-mapping by modeling infinitesimal effects. *Nature genetics*, 56(1):162–169, 2024.
35. Jinyan Teng, Yahui Gao, Hongwei Yin, Zhonghao Bai, Shuli Liu, Haonan Zeng, PigGTEx Consortium, Lijing Bai, Zexi Cai, Bingru Zhao, et al. A compendium of genetic regulatory effects across pig tissues. *Nature genetics*, 56(1):112–123, 2024.
36. David Habier, Rohan L Fernando, and Dorian J Garrick. Genomic blup decoded: a look into the black box of genomic prediction. *Genetics*, 194(3):597–607, 2013.
37. Anna Wolc and Jack CM Dekkers. Application of bayesian genomic prediction methods to genome-wide association analyses. *Genetics Selection Evolution*, 54(1):31, 2022.
38. Shuli Liu, Yahui Gao, Oriol Canela-Xandri, Sheng Wang, Ying Yu, Wentao Cai, Bingjie Li, Ruidong Xiang, Amanda J Chamberlain, Erola Pairo-Castineira, et al. A multi-tissue atlas of regulatory variants in cattle. *Nature genetics*, 54(9):1438–1447, 2022.
39. Zexi Cai, Terhi Iso-Touru, Marie-Pierre Sanchez, Naveen Kadri, Aniek C Bouwman, Praveen Krishna Chitneedi, Iona M MacLeod, Christy J Vander Jagt, Amanda J Chamberlain, Birgit Gredler-Grandl, et al. Meta-analysis of six dairy cattle breeds reveals biologically relevant candidate genes for mastitis resistance. *Genetics Selection Evolution*, 56(1):54, 2024.
40. Wenlong Ma, Yang Fu, Yongzhou Bao, Zhen Wang, Bowen Lei, Weigang Zheng, Chao Wang, and Yuwen Liu. Deepsata: A deep learning-based sequence analyzer incorporating the transcription factor binding affinity to dissect the effects of non-coding genetic variants. *International Journal of Molecular Sciences*, 24(15):12023, 2023.
41. Rongrong Zhao, Rachel Owen, Melissa Marr, J Siddharth, N Chue Hong, Andrea Talenti, Musa A Hassan, and JGD Prendergast. The potential of regulatory variant prediction ai models to improve cattle traits. *bioRxiv*, pages 2024–08, 2024.
42. Dat Thanh Nguyen, Quan Hoang Nguyen, Nguyen Thuy Duong, and Nam S Vo. Lmtag: functional-enrichment and imputation-aware tag snp selection for population-specific genotyping arrays. *Briefings in Bioinformatics*, 23(4):bbac252, 2022.