Screening and assessment of functional variants by regulatory features from epigenomic data in livestock species

Ruixian Ma[1,#], Renzhuo Kuang[1,#], Jingcheng Zhang[2,#], Yueyuan Xu[1], Zheyu Han[1], Mingyang Hu[1], Daoyuan Wang[1], Yu Luan[1], Yuhua Fu[1], Yong Zhang[2], Xinyun Li[1], Mengjin Zhu[1], Tao Xiang[1,*], Shuhong Zhao[1,3,4,*] and Yunxia Zhao[1,4,*]

[1] Key Lab of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education and Key Laboratory of Swine Genetics and Breeding of Ministry of Agriculture, College of Animal Science and Technology, Huazhong Agricultural University, Wuhan, China.2

[2] Key Laboratory of Animal Biotechnology of the Ministry of Agriculture, Northwest A&F University, Yangling, China

[3] Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan 430070, China.

[4] Yazhouwan National Laboratory, 8 huanjin Road, Yazhou District, Sanya City, Hainan Province, 572024, China.

[#] These authors contributed equally

Corresponding author: Tao.Xiang@mail.hzau.edu.cn; shzhao@mail.hzau.edu.cn; yxzhao@mail.hzau.edu.cn

Abstract

Single nucleotide polymorphisms (SNPs) and small insertions/deletions (2-50bp) in genomic regulatory regions may impact function, and although widespread, they are largely unexplored in livestock. Here leveraging >500 epigenomic datasets from pigs, cattle, sheep, and chickens, 8-39 million variants were identified with candidate functional confidence. Using our Functional Confidence scoring system, these candidate functional variants were further ranked as High, Moderate, Low, Minimal, or Possible functional confidence by scoring for likelihood of disrupting transcription factor (TF)-chromatin binding based on their presence in eight genomic regulatory features. Predictive reliability analysis of estimated breeding values (EBVs) based on High/Moderate Confidence variants from pig shows a 23~46% increase in reliability compared to EBVs based on general SNPs, illustrating the versatility of Functional Confidence scoring system for identifying potential functional variants in livestock. Therefore, we developed the Integrated Functional Mutation (IFmut) platform and embed the Functional Confidence scoring system for users to effortlessly navigate through epigenomic data or pinpoint specific genomic features/regions, uncover potential function of new variants or previously identified ones. Our work offers the scientific community a powerful and flexible tool, tailor-made for delving deep into variant function, setting a new benchmark in livestock research and breeding strategies.

Introduction

Variants in cis-regulatory elements can greatly affect gene expression, consequently influencing organismal phenotype (*1-4*).While some variants are harmless (i.e., neutral), others can lead to severe diseases or other deleterious effects (*5*). In human disease research, the identification and characterization of variants are crucial for determining the underlying causes of genetic disorders (*6-8*). Several large-scale initiatives, such as the 1000 Genomes Project, have facilitated exploration of variants in the human genome (*9, 10*). By contrast in livestock research, variants are typically studied through resequencing in livestock populations to improve breeding strategies (*11-14*). The identification of variants is a crucial initial step, but determining whether a variant is functional in livestock species such as pig (Sus scrofa), cattle (Bos taurus), sheep (Ovis aries), and chicken (Gallus gallus) poses a great challenge.

Advances in publicly available bioinformatic analytical toolkits, e.g., the Encyclopedia of DNA Elements (ENCODE) (*15*), have driven considerable progress in functional variant screening in humans, uncovering previously unrecognized functions of several regions in the human genome. In addition, integrated analysis of expression quantitative trait loci (eQTLs) and variants (*16-18*), along with establishment of the regulomeDB database (*19*), has also facilitated identification of cis-regulatory elements and trans-acting factors that influence gene expression, revealing a variety of regulatory mechanisms of the human genome. Despite these innovations in human genomic research, the exploration of variants with regulatory function in the genomes of important livestock species remains limited.

Genomic selection (GS) (*20*) has brought about a revolution in livestock and poultry breeding, enabling greater precision in the selection of individuals based on quantitative traits, such as growth rates or disease resistance (*21-25*). However, GS efficiency relies on SNP markers distributed throughout the genome because all of the QTLs and and the SNPs used for these analyses are in linkage disequilibrium (LD) (*20, 26, 27*). Although GS can enhance the reliability of estimated breeding values (EBVs), EBVs based on GS markers ignore the potentially significant source of functional

genetic variation available in genomic regulatory elements. Variants in regions containing gene regulatory elements can potentially modulate gene expression (*28*). By integrating variants and SNPs in regulatory regions into genomic prediction analysis, the reliability of EBVs for target traits may be increased. However, despite their possible value for improving EBV reliability, studies identifying candidate functional variants that may be informative for GS are lacking.

In this study, we identified genomic regulatory features in over 500 datasets comprising transposase-accessible chromatin with sequencing (ATAC-seq) data, DNase I hypersensitive site sequencing (Dnase-seq) data, H3 lysine 27 acetylation (H3K27ac) ChIP-seq, and transcription factor ChIP-seq data from pigs, cattle, sheep, and chickens. We then identified the candidate functional variants and employed a scoring system to assess the likelihood of variants affecting regulatory function (i.e., Functional Confidence Score) based on their presence (or absence) in eight different regulatory features/regions in genomes of the four livestock species. The identified variants were then ranked into five categories (12 sub-categories) based on Functional Confidence scores, in descending order from High, Moderate, Low, and Minimal functional confidence, to possible association with regulatory function. We further tested whether genomic prediction with High and Moderate confidence IFmut variants identified from three tissues of pig could improve the predictive reliability of EBVs over that of EBVs based on 11000 randomly selected SNP markers in pig. We then constructed the Integrated Functional Mutation (IFmut) database and Functional Confidence scoring system to provide a public resource for researchers. This study provides a large database with a versatile and powerful online toolkit, along with a proof-of-concept demonstration of IFmut for exploration of functional variants in fundamental research and molecular breeding of livestock.

Results

A large scale epigenomic screen of potential functional variants across four livestock species

To screen for functional variants that potentially affect gene expression in livestock, we first obtained SNP and small InDel genomic variants in pig (susScr11), cattle (bosTau9), sheep (oviAri4), and chicken (galGal5) from the Ensembl database, including more than 63 million in pig, over 97 million in cattle, over 63 million in sheep, and over 23 million in chickens. Analysis of their distribution and predicted effects using ChIPseeker R package and SnpEff software indicated that more than 90% of the variants were located in non-coding regions in all four species (Fig. 1A-D; Supplemental Fig. 1), aligning well with previous studies that showed variants are highly prevalent in intergenic and intronic regions of the human genome (*7, 29, 30*). Thus how to identify the potential functional variants from a large pool across four livestock species is still a challenge.

Since genomic regulatory features, especially transcription factor binding sites (TF binding sites) identified by epigenomic analyses, can be informative of the potential function of variants (*19, 31*), we sought to screen for potential function variants in the above libraries using epigenomic data. To this end, we collected 583 total epigenomic datasets (including ATAC-seq, Dnase-seq data, H3K27ac ChIP-seq, TF ChIP-seq and Hi-C) from pigs, cattle, sheep, and chickens generated in previous studies such as the FANNG project (*32, 33*), and our own previous study (*34*). After processing raw reads from ATAC-seq, Dnase-seq, or ChIP-seq data using the ENCODE pipeline, we removed 6 samples due to low number (<10000) of significant peaks and 35 samples due to low correlation among biological replicates (R<0.8). Ultimately, 538 datasets from 19 tissues and 12 cell lines met quality control standards (Fig. 1E,F), including 256 ATAC-seq, 26 Dnase-seq data, 167 H3K27ac ChIP-seq, 80 TF ChIP-seq (63 for CTCF from pigs, cattle, sheep, and chicken and 17 for RAD21, EGR1, KLF2, KLF4,

OSR1, OSR2, SMC2, CAP-H and BRD4 from chicken), and 9 Hi-C datasets from pig (Fig. 1G-J). In total, 125, 193, 24, and 196 total datasets were compiled for pigs, cattle, sheep, and chickens, respectively (Fig. 1E).

ENCODE guidelines (https://www.encodeproject.org/) were then applied to identify genomic regulatory regions containing basic regulatory features and/or TF binding site-related features using these datasets. In total, more than 350000 non-redundant genomic regions with basic regulatory features were identified across all four species, including open chromatin regions (OCR), H3K27ac significant peaks, and nucleosome-free regions (NFR; Table 1). Furthermore, footprint calling and significant TF binding peak calling followed by genomic mapping with TF motif positional weight matrices (PWMs) in the called features yielded between 41460-171290 non-redundant genome regions with regulatory features related to TF binding sites in each species (Table 1). The total length of non-redundant genomic regulatory regions accounted for approximately 31.56% of the pig reference genome (susScr11), while in cattle (bosTau9), sheep (oviAri4), and chicken (galGal5), these accounted respectively for 30.74%, 12.63%, and 37.35% (Table 2).

Genomic variants positioned within transcription factor binding sites, such as in RegulomeDB, often result in functional consequences (*19*). Since the above genomic regions containing basic regulatory and TF binding-related features were identified through DNA-TF interaction data, variants detected in these regions were likely to have transcription regulation function in the host livestock species. Using BEDTools, we then determined which variants in our initial calling were located in these regulatory features, which yielded 21005715 (32.90%; Fig. 2A) SNPs and small InDels in pig, while 39157953 (40.32%; Fig. 2B) were detected in the cattle genome, 8194045 (12.97%; Fig. 2C) in sheep, and 10896983 (47.05%; Fig. 2D) in chicken, which we collectively designated as potential functional variants.

A scoring system to rank variants by likelihood of functional impacts

In the current study, we identify a multitude of variants with predicted functional/phenotypic consequences in four livestock species, and the number of such variants in each species was positively correlated with proportion of the genome occupied by regulatory features (R=0.74; Supplemental Fig. 1E). To further distinguish differences in the likelihood that a regulatory region variant will indeed impact transcription regulation in livestock species, we developed a functional confidence index similar to that used by RegulomeDB for variant classification with human TF ChIP-seq data (*19*). At present, only 80 TF ChIP-seq datasets are available in livestock, the vast majority of which were generated for CTCF (in total 63), with only chicken having 17 ChIP-seq data for 9 TFs with ChIP-seq data, compared to the 876 TFs covered by 3537 ChIP-seq data in RegulomeDB v.2. Thus, due to the lack of TF ChIP-seq data in livestock, functional confidence scoring instead relied on a combination of ATAC-seq/Dnase-seq (i.e., OCR and footprints) and H3K27ac ChIP-seq (i.e., NFR and significant narrow peaks; Fig3A; Table 3). In addition, quantitative trait loci (QTL) data were also collected, since variants in these regions can also potentially impact agronomic traits (Supplemental Table 1).

In the Functional Confidence scoring system, the greater the number of regulatory features used to determine the presence of SNPs/small InDels in TF binding sites, the higher the likelihood that a variant could affect transcriptional regulation (Table 3). Based on the prominent association of NFRs, OCRs and TF footprints (especially those containing fully or partially matching recognition motifs) with transcriptional activation, variants in these regions had the highest likelihood of affecting TF binding and gene expression, and were therefore scored as high functional confidence variants (Category 1). Variants that met these criteria but were never found in NFRs were subsequently scored as moderate functional confidence (Categories 2a-2d), suggesting a moderate likelihood of affecting TF activity. Moreover, within Category 1 and 2, variants were present in QTLs, were assigned higher scores

(Categories 1a, 1c, 2a and 2c respectively), whereas variants were not in QTLs, had slightly lower likelihoods (Categories 1b, 1d, 2b and 2d respectively). By contrast, variants found in DNase/ATAC-seq or H3K27ac ChIP-seq data but not in TF footprints or recognition motifs were included in Categories 3 and 4, with low functional confidence and minimal functional confidence, respectively, in their likelihood of affecting TF binding. Finally, Category 5 was reserved for variants detected only by H3K27ac ChIP-seq, and were therefore potentially associated with transcriptional regulation (see Table 3 for a key of criteria).

Next, the candidate functional variants identified by our study were ranked based on our Functional Confidence scoring system. Then, Figure 3 shows a summary of variant numbers in each functional confidence category for pig (Fig.3E), cattle (Fig. 3B), sheep (Fig. 3C), and chicken (Fig. 3D). Among these variants, a total of 240938, 3096314, 280103, and 204100 SNPs/small InDels were included in high and moderate functional confidence categories (Category 1 and 2) Categories 1 and 2, accounting for 1.15%, 7.91%, 3.42% and 1.87% of all potential functional variants in pigs, cattle, sheep and chickens, respectively (Fig. 3B-E).

To validate the variants in our above analysis were present in population data and that functional confidence scoring could be applied to whole-genome sequencing (WGS) data, we obtained 22926176 minimum allele frequency (MAF>0.047) filtered variants in WGS data from 491 individual pigs across 61 breeds generated in our previous study (*34*). Among these variants, 7557763 (32.97%) were identified as potentially functional variants, 87002 (1.15%) of which fell into categories 1 or 2 (Fig. 3F). Overall, the proportions of variants in each category filtered by MAF from WGS data were similar to that of variants obtained from Ensembl (Fig. 3B,F). These results indicated that taking MAF into account did not affect the proportion of variants in each category, but could reduce the number of candidate functional variants. Thus on animal breeding a lower MAF threshold (e.g. 0.01) have to consider for functional variants to keep their efficiency of animal breeding.

The functional confidence scoring in eQTL classification and EBV reliability assessment

Although eQTLs are reportedly associated with gene expression (*35, 36*), some TF ChIP-seq and DNase-seq studies in humans suggest that more than 50% of eQTLs are not associated with TF binding sites (*19, 37*), implying that genomic regulatory features could be used to assess the potential regulatory function of eQTLs. To test this possible use of our Functional Confidence scoring system, we obtained cis-eQTL data from adipose, liver, spleen, hypothalamus, kidney, lung, muscle, and rumen of cattle from the farmGTEx database (https://www.farmgtex.org/). Among these cis-eQTLs, more than 58% had no classification as potential functional variants (Categories 1-5; Fig. 4A and Supplemental Table 2). Moreover, only a small fraction of cis-eQTLs (~2.60%) in each tissue were scored as high and moderate functional confidence variants (Category 1 and 2 variants; Fig. 4B,C). These results indicated that cis-eQTLs could primarily serve as marker loci, but were unlikely to be functional variants that affect transcription. In addition, this analysis provided a proof-of-concept that Functional Confidence scoring system could be used to assess potential regulatory function in cis-eQTL datasets and score for functional confidence.

We further validated Functional Confidence scoring system variant identification and functional confidence scoring by genomic prediction with high and moderate functional confidence variants (Category 1 and 2 variants) in pigs. We assessed the predictive reliability of estimated breeding values (EBVs) for two traits, average daily gain (ADG) and backfat thickness (BF) in a large white population (n=874) using a genomic BLUP model with DMU software (*38*). EBVs were based on four different genomic relationship matrices constructed by four scenarios of SNP markers, including three scenarios using high and moderate functional confidence variants from muscle, liver, or adipose, as well as one scenario that used 11k randomly selected variants from whole genome sequence of pig (Table 4). Overall, the predictive reliability of EBVs for ADG and trait BF was similar among the three scenarios using high and moderate functional confidence variants (~0.31-0.38), whereas the predictive

reliability of EBVs was lowest in the scenario using 11K random SNPs (~0.27), despite containing the highest number of SNP markers (11000). Notably, predictive reliability was highest in the scenario based on variants detected in adipose (~0.38), despite using the fewest markers (3861 SNPs). Predictive reliability of EBVs generated with Category 1 and 2 variants from liver was higher than that of functional confidence variants from muscle. Ultimately, the predictive reliability of EBVs increased 23%~46% for the three tissue types by using high and moderate functional confidence variants compared to EBVs based on randomly selected SNP markers. This analysis further validated the use of Functional Confidence scoring system for screening functional variants in genomic data of livestock.

Development of the Integrated Functional Mutation database for screening candidate functional variants in livestock species

In order to facilitate screening for candidate functional variants in livestock species, we integrated genomic variants with epigenomic datasets in a single database, the Integrated Functional Mutation (IFmut) database. This database contains 65124531 potential functional variants from the genomes of pig, cattle, sheep, and chicken (Fig. 5A), as well as the 538 aforementioned epigenomic datasets from 19 tissues and 12 cell lines across the four species (Fig. 5B). In addition, the IFmut database (http://www.ifmutants.com:8210/#/home) has a user-friendly web interface that enables users to query variants of interest, different genomic regions, or browse epigenomic signal viewers.

In the first module, users can use a "Quick Search" function on the homepage to search for a specific dbSNP ID or search specific genomic regions for a variant of interest. Details, such as genomic location, conversion type, and functional confidence score (defined in the following section) about the queried variant, if stored in the IFmut database, are then listed at the bottom of the homepage (Fig. 5C). Clicking on an SNVID in the search hits will direct the user to a new page containing information about the regulatory feature(s) associated with queried variant of interest (Fig. 5D). In the

third module, users can search for “Affected motif” to facilitate hypothesis generation about the potential effects of a variant on TF binding. Searches in this module return logos plots of conservation of the potentially affected TF motif(s) and a table containing the predicted effect on TF binding, and the affected gene symbol of the TF motif etc. (Fig. 5E,F). In the fourth module, the "JBrowse" features allows users to view ATAC-seq, Dnase-seq and ChIP-seq (H3K37ac and TFs) signals or Hi-C interaction heatmaps (for Pig) around the variant of interest, as well as nearby genes (Fig. 5G). For this purpose, each epigenomic dataset in the IFmut database is accompanied by BigWig and genome annotation files that can be loaded in the right sidebar of JBrowse (Fig. 5G; Supplemental Fig. 2), allowing users to examine epigenomic signals or annotation data around queried variants in greater detail.

To facilitate further exploration of potential functional variants, IFmut also provides hyperlinks to other databases: (i) For variants in pigs and cattle, users can click on hyperlinked SNVIDs to access the IAnimal database (https://ianimal.pro/), which contains additional information, such as genotype and major allele frequency (Fig. 5H). (ii) Clicking on the “TAD/TAD Boundary” feature of IFmut entries that contain topologically associating domain (TAD) information related to genomic variants in pig will also direct users to the IAnimal database, allowing a subsequent search for genes within that TAD or TAD boundary (Fig. 5I). (iii) Since ChromHMM Chromatin States uses epigenomic information (such as ChIP-Seq data for various histone modifications) across one or more human cell types to facilitate annotation of non-coding genome regions, this function can be used for comparative genomics analysis to identify regulatory feature-containing regions. The "ChromHMM Chromatin States" section in the IFmut database can thus be used to map variant-containing genomic regulatory regions in the four livestock species to corresponding chromatin regions in the human hg38 genome (https://genome-asia.ucsc.edu/; Fig. 5J) by LiftOver (*39*).

It should be noted that IFmut also incorporates the details of functional confidence scoring for each variant and provides the tool for scoring novel variants. For such variants that are not yet included in IFmut, and are the subject of a user query, a dialog box will prompt the user to categorize their variant using an embedded “Variant

scoring tool" (Fig. 6A). Upon clicking the "OK" button, a window is displayed containing the classification results for the variant of interest (Fig. 6A). Queried SNPs can also be loaded into JBrowse to visualize the relevant epigenetic data (Fig. 6B).

Discussion

Previous studies have shown that the majority of SNPs and small InDels are located in non-protein-coding genomic regions (*7, 40-44*), and thus interpreting whether and how a variant may affect function remains considerably challenging (*45-47*). Evaluating perturbation effects of variants on TF binding sites in TF ChIP-seq data is a demonstrably effective way for identifying potential functional variants in the human genome (*19*). However, available TF ChIP-seq data is still comparatively lacking in livestock, posing an obstacle for this approach of screening functional variants in regulatory genomic features. To overcome this limitation, we compiled the IFmut database of candidate SNP and small InDel functional variants in or near TF binding sites in ATAC-seq/Dnase-seq and H3K27ac ChIP-seq datasets.

ATAC-seq/Dnase-seq analyses can largely capture TF binding footprints in full range of open chromatin regions across the genome, and have been widely used for this purpose in human and livestock research (*48-52*). At present, the TF binding sites capturing in livestock were primarily relay on the ATAC-seq/Dnase-seq rather than the TF ChIP-seq. Then our Functional Confidence scoring system used ATAC-seq/Dnase-seq data to identify TF binding sites, which is different with the approach in RegulomeDB based on TF ChIP-seq (*19*). Overall, our scoring approach was more suitable the current study of functional mutations in livestock for abundant ATAC-seq/Dnase-seq datasets in these species, as well as the design idea of using the ATAC-seq/Dnase-seq data to identify TF binding sites to rank functional mutations can also be transplanted to related research works on other species.

Further, our scoring approach identified five main categories of candidate functional variants in pig, cattle, sheep, and chicken. We primarily focused on SNP and InDel variants in the High and Moderate confidence groups (Categories 1 and 2), since

these variants were ranked based on their elevated likelihood of producing an effect in livestock breeding and production. Genomic predictions with these High and Moderate variants showed that Functional Confidence scoring could increase the predictive reliability of EBVs in pigs compared to a larger set of randomly selected SNPs. These findings suggested that our scoring system can guide the identification of important variants, and could therefore drive advances in genetic improvement of livestock.

It is well-known that increasing the number of SNP markers can also increase the predictive reliability of GEBVs (genomic EBVs (*20*)). Nevertheless, in this study, we found that genomic prediction with 11000 random SNPs from across the pig genome resulted in markedly lower GEBV reliability than that in some scenarios where even only one third the number of High/Moderate confidence SNPs from IFmut were used. Furthermore, this genomic prediction analysis also indicated that adipose tissue was more strongly associated with average daily gain and backfat thickness than muscle or liver. This finding might be at least partially explained by adipose function as the major site of energy storage and insulation in pigs (*53*), and provides direct evidence that the selection of candidate functional SNPs can guide genomic breeding efforts in pigs.

As variants play important roles in genomic breeding in livestock, a number of sequencing-related databases have been developed for animal research, such as AnimalQTLdb (*54*), Animal-ImputeDB (*55*), Animal-eRNAdb (*56*), and IAnimal (*57*), and range from one omics data type to comprehensive multi-omics data collections. However, tools for identifying candidate functional variants, visualizing relevant evidence in epigenetics data, and scoring for confidence in their function are still unavailable for mining these databases. We therefore designed the IFmut platform to allow users to retrieve and explore genomic, and epigenetic data related to the possible function of a variant, as well as a Functional Confidence scoring tool for assessing new variants of interest identified by users alongside those in IFmut and across multiple livestock species. Overall, the Functional Confidence classification data for SNPs and small InDels in the four species in IFmut, along with the tools for further exploration, can facilitate investigations of functional impacts of variants

Methods

Data collection

Genome variants VCF data of pig (susScr11), cattle (bosTau9), and chicken (galGal5) were downloaded from ensemble database (http://ftp.ensembl.org/pub/), genome variants VCF data of sheep (oviAri4) was from NCBI Single Nucleotide Polymorphism Database (https://ftp.ncbi.nih.gov/snp/organisms/archive/sheep_9940/VCF/00-All.VCF.gz). We also used whole-genome sequencing (WGS) data from 491 individual pigs across 61 breeds generated in our previous study(*29*). QTL data of four livestock were downloaded from Animal QTL database (https://www.animalgenome.org/cgi-bin/QTLdb/). In cattle, we also downloaded the best variants cis-eQTL data from the farmGTEx database (https://cgtex.roslin.ed.ac.uk/wp-content/plugins/cgtex/static/rawdata/Full_summary_statisitcs_cis_eQTLs_FarmGTEx_cattle_V0.tar.gz). The TF ChIP-seq, H3K27ac ChIP-seq, and ATAC-seq data in pig, cattle, sheep, and chicken, Hi-C data in pig, and Dnase-seq in chicken were downloaded from NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra/). A total of 579 raw epigenomic datasets were collected from multiple projects in NCBI， of which 75 datasets from pig were from our previous study (*34*).

Sequencing data analysis

To adhere to the ENCODE standard, we primarily refer to the analysis methods used in our previous study for processing ChIP-seq and ATAC-seq data (*34*).

ChIP-seq

Mapping and Quality control

The ENCODE ChIP-seq pipeline (https://github.com/kundajelab/chipseq_pipeline) was utilized to process the ChIP-seq datasets of the four species in a strict manner. The raw reads from each dataset were aligned to the respective reference genome assemblies (susScr11, bosTau9, oviAri4, galGal5) using BWA v0.7.17 (*58*). Subsequently, the removal of low MAPQ reads (<25), unmapped reads, mate unmapped reads, not

primary alignment reads, and duplicate reads using Picard v1.126 (https://broadinstitute.github.io/picard) and SAMTools v1.9 (*59*).

The read coverage of genomic regions between replicate filtered BAM files was computed using the multiBamSummary bins function of deepTools v2.0 (*60*). A bin size of 2 kb was used to assess genome-wide similarities. The resulting read coverage matrix obtained from the multiBamSummary step was used to calculate the Pearson correlation coefficients between two replicate filtered BAM files. The non-duplicated BAM file of replicates with a Pearson correlation coefficient ≥ 0.8 were merged, and the remaining replicates with a correlation coefficient < 0.8 were excluded from further analysis.

Identification of nucleosome free region

The HOMER (*61*) were utilized to detect nucleosome-free regions (NFR). The makeTagDirectory command was used to generate tag directories for the H3K27ac IP and input data using the merged non-duplicated BAM file obtained from the "Mapping and Quality control" steps. Subsequently, the findPeaks command with the -nfr option was applied to identify NFR peaks, requiring at least 10,000 peaks per data, and finally excluding the scaffold regions.

Identification of TF binding sites and H3K27ac narrow peaks

The identification of TF binding sites and H3K27ac narrow peaks was carried out using MACS2 v2.1.0 (*62*) and deepTools v2.0 (*60*), as described in greater detail in the methods section of our previous study (*34*).

ATAC-seq

Mapping, quality control and peak calling

The ATAC-seq datasets of four species were processed following the ENCODE ATAC-seq pipeline (https://github.com/kundajelab/atac_dnase_pipelines). The preprocessing steps included checking and trimming adapters using Cutadapt v1.14 (https://cutadapt.readthedocs.io/en/stable/). The ATAC-seq reads were then aligned to the susScr11, bosTau9, oviAri4, and galGal5 reference genome assemblies using

Bowtie2 v2.3.4.1. After alignment, low MAPQ reads (<25), unmapped reads, mate unmapped reads, not primary alignments, reads failing platform, and duplicates were removed using SAMTools v1.9 (*59*) and Picard v1.126 (https://broadinstitute.github.io/picard) software. The mitochondrial reads were further removed from the mapped BMA file using BEDTools v2.26.0 (*63*) to generate effective reads, which were subsequently used for peak calling. MACS2 v2.1.0 (*62*) was employed to call peaks for each replicate individually, using parameters: genome size (-g), p-value threshold (0.01), peak model (--nomodel), shift size (--shift), extension size (--extsize), and other options (--B, --SPMR, --keep-dup all, --call-summits). And generate a data set of at least 10,000 peaks for further analysis.

Dnase-seq

Mapping, quality control and peak calling

For the Dnase-seq datasets of chicken, the ENCODE Dnase-seq pipeline (https://github.com/kundajelab/atac_dnase_pipelines) was followed. With the 'dnase_seq' parameter specified to indicate Dnase-seq data, and the others were consistent with the above ATAC-seq analysis.

Identification of open chromatin region

In the peak calling step, peaks with $P < 10^{-5}$ were considered significant and selected for further analysis. These significant narrow peaks were filtered based on replicates with high Pearson correlation coefficients (R > 0.8). The peaks from these replicates were merged using BEDTools v2.26.0 (*63*), requiring at least 50% overlap between peaks in each replicate. The merged peaks represent open chromatin regions. Furthermore, the BAM files from highly correlated replicates (R > 0.8) were merged to generate signal tracks using MACS2 v2.1.0 (*62*). This step helps to visualize the signal intensity and distribution of chromatin accessibility across the genome.

Identification of footprints in ATAC-seq and Dnase-seq

The footprint analysis was primarily performed as the following steps: (i) the board peaks were called from the merged ATAC-seq or Dnase-seq data using the MACS2

v2.1.0 broad module (*62, 64*); (ii) the broad peaks meeting the criteria of $P < 10^{-10}$ and $10^{-10} < P < 10^{-5}$ overlapping OCR were merged with BEDTools v2.26.0 (*63*) as significant broad peaks; (iii) the Hmm-based IdeNtification of Transcription factor footprints (HINT) framework of Regulatory Genomics Toolbox (RGT) v0.13.2 (*65*) was employed to analysis footprints using the significant broad peaks. The HINT framework was utilized with specific parameters depending on whether ATAC-seq or Dnase-seq data was used (--atac-seq or --dnase-seq) and considering paired-end sequencing data (--paired-end). The organism information (--organism=) was also specified; and (iv) the cutoff value for footprint score was determined as more than the 20% quantile of all footprint score generated by the HINT framework of GRT v0.13.2 (*65*).

Transcription factor motif mapping in genome function region

The transcription factor motif mapping was primarily performed as the following steps: (i) OCR, NFR, TF binding sites and footprint in OCR regions were merged into a BED file; (ii) The fasta-get-markov command from the MEME Suite (https://github.com/cinquin/MEME) software was used to generate a .fa.bg file and "bedtools getfasta" command generate .fa file corresponding .bed file of step (i); (iii) The fimo command in MEME Suite (--max-stored-scores 5000000) used to map motif in the genome; and (iv) the fimo mapped results of pig, cattle, sheep filtered by $P < 5*10^{-6}$, and chicken filtered by Pvalue $< 5*10^{-7}$.

Prediction of transcription factor motif effects

In addition, potential functional variants (Categories 1-5) located within footprint regions were analyzed using the motifbreakR (*66*) package in R v4.0. The motifDB database, specifically JASPAR 2018 (*67*), was selected as the data source for predicting the transcription factors to which the SNPs may bind.

Hi-C

The Hi-C data of two-week-old LW pigs were from our previous study (*34*), and the other Hi-C data were downloaded from GEO under accession number GSE153452 at

http://ncbi.nlm.nih.gov/geo, including the cells of pig from zygotes, 4 cell stage and morula of in vitro fertilization (IVF), and pig embryonic fibroblasts (PEFs). These downloaded data were processed using the HiC-Pro (version 2.11.1) pipeline to produce the ICE normalization contact matrices (*68*). The insulation score of the ICE matrix was calculated by using the following options: -is 480000 -ids 320000 -im iqrMean -ss 160000. Furthermore, the insulation method was utilized to define the topologically associating domain (TAD) structure (insulation/boundaries).

Variants distribution statistics

SNPs and small InDels were annotated using ChIPseeker package in R v3.6.0, the parameter of annotatePeak was setted that including level="transcript", assignGenomicAnnotation=TRUE, genomicAnnotationPriority=c("Promoter", "5UTR", "3UTR", "Exon", "Intron","Downstream", "Intergenic"), annoDb=NULL, addFlankGeneInfo=FALSE,sameStrand=FALSE,ignoreOverlap=FALSE, ignoreUpstream=FALSE,ignoreDownstream=FALSE). Next, the reference genome file (fasta) and annotation files (gtf) were used with the snpEff v4.5 software to predict the effects of SNPs on known genes (java -Xmx8g -jar snpEff.jar genome -i .bed).

The identification and filtering of pig variants.

Identification of SNP and small InDel in pig

A total of 491 whole-genome sequences from 61 pig breeds were obtained from our previous study (*29*). The method of data processing was consistent with our previous article (*34*).

Chromatin state discovery and characterization

The chromatin states of human genome (hg38) were downloaded from the NIH Roadmap Epigenomics program (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/). The genome coordinates of human genome chromatin states were converted into those of pig genome (SusScr11), chicken genome (GalGal5), cattle

genome (BosTau9), and sheep genome (OviAri4) by LiftOver, respectively. The positions of SNPs from pig, chicken, cattle, and sheep were used to overlap with the converted genome coordinates of chromatin states by BEDTools v2.26.0. In addition, the converted genome coordinates of chromatin states with the same SNP were merged, and the merged genome coordinates were transformed into those of human genome.

Performances of genomic predictions

Dataset

The phenotypic dataset used for genomic prediction were obtained from a national pig nucleus herd in North China. In this study, we used phenotypic recordings for two productive traits: 30-100 kg average daily gain (ADG) and 100 kg backfat thickness (BF). All the phenotypic records for the traits were obtained at the same time point, allowing a 10-kg deviation from the final bodyweight (100 ± 10 kg). All of the phenotypes were recorded between the year early 2018 and October 2022. Based on the traced pedigree, there were 11 lines existing in such pig population. For each pig line, DNA samples were collected from about 80 distantly related pigs and were sequenced by DNBSEQ-T7 platform with an averaged 5× coverage. In total, 874 pigs were sequenced. After quality controls, which includes a genotype missing rate below 10%, a call rate of SNPs above 90%, and a minimum allele frequency (MAF) above 1%, 18460807 (18000K) SNPs were kept and analyzed in the following study. Missing genotypes were imputed using software Beagle version 5.3. Among the 874 sequenced pigs, 872 pigs had phenotypes of ADG data, meanwhile 867 pigs had BF recordings. Environmental factors including such as genders, herds, and physical units were completely recorded.

Genomic Best Linear Unbiased Prediction (GLUP) models

The breeding values (EBV) for different traits were estimated using the following GBLUP models:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{e},$$

where $\boldsymbol{y}$ represents a column vector of phenotypic values for each trait; b represents a

vector of fixed effects, including sex effect, herd effect and physical units effects; $\boldsymbol{u}$ represents a vector of random additive genetic effects; e represents a vector of residual effects. Matrices X and Z are corresponding design matrices associated with these effects. The GBLUP model assumes a normal distribution for the random additive effects and residual effects, as $\boldsymbol{u} \sim N(0, \boldsymbol{G}\sigma_u^2)$, where $\boldsymbol{G}$ is genomic relationship matrix constructed as Vanraden method 1; $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$, where $\boldsymbol{I}$ is an identity matrix. The additive genetic variance and residual variance are denoted by σ_u^2 and σ_e^2, respectively.

Scenarios of constructing genomic relationship matrices

In this study, GBLUP models with four different genomic relationship matrices ($\boldsymbol{G}$) were used to estimate the GEBVs for both ADG and BF traits. Four different sets of SNP markers were used for constructing the corresponding $\boldsymbol{G}$ matrix. In scenario 1, sequenced SNP markers that were with top 1 and top 2 muscle scores (1+2 muscle, 10544 SNPs) were calculated the $\boldsymbol{G}$ matrix. Similarly, sequenced SNP markers that were with top 1 and top 2 liver scores (1+2 liver, 6049 SNPs) and with top 1 and top 2 adipose scores (1+2 adipose, 3801 SNPs) were used for constructing $\boldsymbol{G}$ matrices in scenarios 2 and 3, respectively. In scenario 4, 11000 (11K) randomly selected SNP markers were used for constructing $\boldsymbol{G}$ matrix. Scenario 4 were repeated for three times in the study.

Predictive Reliabilities

The mean predictive reliabilities of GEBVs were determined by employing the subsequent formula (Mrode, 2005):

$$r^2 = \sum(1 - \frac{SEP_i^2}{\sigma_g^2}) / N,$$

where r^2 is reliability of GEBVs and i denotes an individual animal i; SEP represents the standard error that is associated with the predicted GEBVs; σ_g^2 represents the additive genetic variance and N is the number of used animals.

Data access

All track of ATAC-seq, ChIP-seq (H3K27ac and TFs) and Hi-C, as well as the candidate functional variants and their Functional Confidence score are available at

http://www.ifmutants.com:8210/#/home.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Dr. Christopher K Tuggle (cktuggle@iastate.edu) at Department of Animal Science, Iowa State University for valuable suggestions to our work. This work was partially supported by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (Grant No.32221005), the National Natural Science Foundation of China (35410676), the National Key Research and Development Project (2019YFE0115400) and the Fund of Modern Industrial Technology System of Pig (CARS-35).

References

1. F. Zhang, W. Gu, M. E. Hurles, J. R. Lupski, Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**, 451-481 (2009).
2. J. Wang *et al.*, HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res* **47**, D106-D112 (2019).
3. Z. Pan *et al.*, Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nat Commun* **12**, 5848 (2021).
4. R. Redon *et al.*, Global variation in copy number in the human genome. *Nature* **444**, 444-454 (2006).
5. H. O. Heyne *et al.*, Mono- and biallelic variant effects on disease at biobank scale. *Nature* **613**, 519-+ (2023).
6. T. J. Cherry *et al.*, Mapping the cis-regulatory architecture of the human retina reveals noncoding genetic variation in disease. *Proc Natl Acad Sci U S A* **117**,

9001-9012 (2020).

7. F. Zhang, J. R. Lupski, Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102-110 (2015).

8. S. Nik-Zainal *et al.*, Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54 (2016).

9. O. Devuyst, The 1000 Genomes Project: Welcome to a New World. *Perit Dial Int* **35**, 676-677 (2015).

10. F. S. Collins, L. Fink, The Human Genome Project. *Alcohol Health Res World* **19**, 190-195 (1995).

11. H. Ai *et al.*, Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat Genet* **47**, 217-225 (2015).

12. H. D. Daetwyler *et al.*, Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46**, 858-865 (2014).

13. X. Li *et al.*, Whole-genome resequencing of wild and domestic sheep identifies genes associated with morphological and agronomic traits. *Nat Commun* **11**, 2815 (2020).

14. C. J. Rubin *et al.*, Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587-591 (2010).

15. E. P. Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640 (2004).

16. B. Zeng *et al.*, Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. *Nat Genet* **54**, 161-169 (2022).

17. L. Franke, R. C. Jansen, eQTL analysis in humans. *Methods Mol Biol* **573**, 311-328 (2009).

18. A. C. Nica, E. T. Dermitzakis, Expression quantitative trait loci: present and future. *Philos T R Soc B* **368**, (2013).

19. A. P. Boyle *et al.*, Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-1797 (2012).

20. T. H. E. Meuwissen, B. J. Hayes, M. E. Goddard, Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-1829 (2001).

21. M. P. Calus, Genomic breeding value prediction: methods and procedures. *Animal* **4**, 157-164 (2010).

22. L. F. Brito *et al.*, Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. *BMC Genet* **18**, 7 (2017).

23. X. Ma *et al.*, Prediction of breeding values for group-recorded traits including genomic information and an individually recorded correlated trait. *Heredity (Edinb)* **126**, 206-217 (2021).

24. M. Cappelloni, M. Gallo, A. Cesarani, Use of threshold and linear models to estimate variance components and breeding values for disease resistance in Italian heavy pigs. *Italian Journal of Animal Science* **21**, 488-492 (2022).

25. D. A. L. Lourenco *et al.*, Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. *Genetics Selection Evolution* **47**, (2015).

26. M. W. Bruford, D. G. Bradley, G. Luikart, DNA markers reveal the complexity of livestock domestication. *Nat Rev Genet* **4**, 900-910 (2003).

27. M. L. Whitfield, L. K. George, G. D. Grant, C. M. Perou, Common markers of proliferation. *Nat Rev Cancer* **6**, 99-106 (2006).

28. M. S. Hill, P. Vande Zande, P. J. Wittkopp, Molecular and evolutionary processes generating variation in gene expression. *Nature Reviews Genetics* **22**, 203-215 (2021).

29. Y. Fu *et al.*, A gene prioritization method based on a swine multi-omics knowledgebase and a deep learning model. *Commun Biol* **3**, 502 (2020).

30. E. E. Eichler, Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *N Engl J Med* **381**, 64-74 (2019).

31. T. I. Lee, R. A. Young, Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237-1251 (2013).

624 32. C. Kern, Y. Wang, X. Xu, Z. Pan, H. Zhou, Functional annotations of three
625 domestic animal genomes provide vital resources for comparative and
626 agricultural research. *Nature Communications*.
627 33. S. Foissac, S. Djebali, K. Munyard, N. Vialaneix, E. Giuffra, Multi-species
628 annotation of transcriptome and chromatin structure in domesticated animals.
629 *BMC Biology* **17**, 108 (2019).
630 34. Y. Zhao *et al.*, A compendium and comparative epigenomics analysis of cis-
631 regulatory elements in the pig genome. *Nat Commun* **12**, 2217 (2021).
632 35. R. Joehanes *et al.*, Integrated genome-wide analysis of expression quantitative
633 trait loci aids interpretation of genomic association studies. *Genome biology*
634 **18**, 16 (2017).
635 36. B. D. Umans, A. Battle, Y. Gilad, Where Are the Disease-Associated eQTLs?
636 *Trends in Genetics* **37**, (2020).
637 37. A. E. Handel, G. Gallone, M. Zameel Cader, C. P. Ponting, Most brain
638 disease-associated and eQTL haplotypes are not located within transcription
639 factor DNase-seq footprints in brain. *Human Molecular Genetics* **26**, 79-89
640 (2017).
641 38. J. J. Per Madsen, A user's guide to DMU. Version 6, release 5.2. (2013).
642 39. R. M. Kuhn, H. David, K. W. James, The UCSC genome browser and
643 associated tools. *Briefings in Bioinformatics*, 144-161 (2013).
644 40. H. Giral, U. Landmesser, A. Kratzer, Into the Wild: GWAS Exploration of
645 Non-coding RNAs. *Frontiers in Cardiovascular Medicine* **5**, (2018).
646 41. N. Okumura *et al.*, Genetic relationship amongst the major non-coding regions
647 of mitochondrial DNAs in wild boars and several breeds of domesticated pigs.
648 *Animal Genetics* **32**, 139-147 (2001).
649 42. M. Spielmann, S. Mundlos, Looking beyond the genes: the role of non-coding
650 variants in human disease. *Hum Mol Genet* **25**, R157-R165 (2016).
651 43. Y. Zhu, C. Tazearslan, Y. Suh, Challenges and progress in interpretation of
652 non-coding genetic variants associated with human disease. *Exp Biol Med*
653 *(Maywood)* **242**, 1325-1334 (2017).

654 44. L. D. Ward, M. Kellis, Interpreting noncoding genetic variation in complex
655 traits and human disease. *Nat Biotechnol* **30**, 1095-1106 (2012).
656 45. J. C. Cohen *et al.*, Multiple rare variants in NPC1L1 associated with reduced
657 sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad*
658 *Sci U S A* **103**, 1810-1815 (2006).
659 46. J. P. Hugot *et al.*, Association of NOD2 leucine-rich repeat variants with
660 susceptibility to Crohn's disease. *Nature* **411**, 599-603 (2001).
661 47. A. H. Mirza, S. Kaur, C. A. Brorsson, F. Pociot, Effects of GWAS-Associated
662 Genetic Variants on lncRNAs within IBD and T1D Candidate Loci. *Plos One*
663 **9**, (2014).
664 48. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf,
665 Transposition of native chromatin for fast and sensitive epigenomic profiling
666 of open chromatin, DNA-binding proteins and nucleosome position. *Nature*
667 *Methods* **10**, 1213-+ (2013).
668 49. M. M. Halstead *et al.*, A comparative analysis of chromatin accessibility in
669 cattle, pig, and mouse tissues. *Bmc Genomics* **21**, (2020).
670 50. A. P. Boyle *et al.*, High-resolution mapping and characterization of open
671 chromatin across the genome. *Cell* **132**, 311-322 (2008).
672 51. R. E. Thurman *et al.*, The accessible chromatin landscape of the human
673 genome. *Nature* **489**, 75-82 (2012).
674 52. L. Song, G. E. Crawford, DNase-seq: a high-resolution technique for mapping
675 active gene regulatory elements across the genome from mammalian cells.
676 *Cold Spring Harbor Protocols* **2010**, pdb. prot5384 (2010).
677 53. M. Kojima *et al.*, Differences in gene expression profiles for subcutaneous
678 adipose, liver, and skeletal muscle tissues between Meishan and Landrace pigs
679 with different backfat thicknesses. *Plos One* **13**, (2018).
680 54. Z. L. Hu, C. A. Park, J. M. Reecy, Bringing the Animal QTLdb and CorrDB
681 into the future: meeting new challenges and providing updated services.
682 *Nucleic Acids Research* **50**, D956-D961 (2022).
683 55. W. Q. Yang *et al.*, Animal-ImputeDB: a comprehensive database with multiple

684 animal reference panels for genotype imputation. *Nucleic Acids Research* **48**,
685 D659-D667 (2020).
686 56. W. W. Jin *et al.*, Animal-eRNAdb: a comprehensive animal enhancer RNA
687 database. *Nucleic Acids Research* **50**, D46-D53 (2022).
688 57. Y. H. Fu *et al.*, IAnimal: a cross-species omics knowledgebase for animals.
689 *Nucleic Acids Research*, (2022).
690 58. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-
691 Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
692 59. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools.
693 *Bioinformatics* **25**, 2078-2079 (2009).
694 60. F. Ramirez, F. Dundar, S. Diehl, B. A. Gruning, T. Manke, deepTools: a
695 flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**,
696 W187-191 (2014).
697 61. S. Heinz *et al.*, Simple combinations of lineage-determining transcription
698 factors prime cis-regulatory elements required for macrophage and B cell
699 identities. *Mol Cell* **38**, 576-589 (2010).
700 62. Y. Zhang *et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**,
701 R137 (2008).
702 63. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing
703 genomic features. *Bioinformatics* **26**, 841-842 (2010).
704 64. T. Liu, Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads
705 generated by sequencing protein-DNA interactions in embryonic stem cells.
706 *Methods Mol Biol* **1150**, 81-95 (2014).
707 65. Z. Li *et al.*, RGT: a toolbox for the integrative analysis of high throughput
708 regulatory genomics data. *BMC Bioinformatics* **24**, 79 (2023).
709 66. S. G. Coetzee, G. A. Coetzee, D. J. Hazelett, motifbreakR: an R/Bioconductor
710 package for predicting variant effects at transcription factor binding sites.
711 *Bioinformatics* **31**, 3847-3849 (2015).
712 67. A. Khan *et al.*, JASPAR 2018: update of the open-access database of
713 transcription factor binding profiles and its web framework. *Nucleic Acids Res*

714 **46**, D1284 (2018).
715 68. N. Servant *et al.*, HiC-Pro: an optimized and flexible pipeline for Hi-C data
716 processing. *Genome Biol* **16**, 259 (2015).
717

Figures and Tables

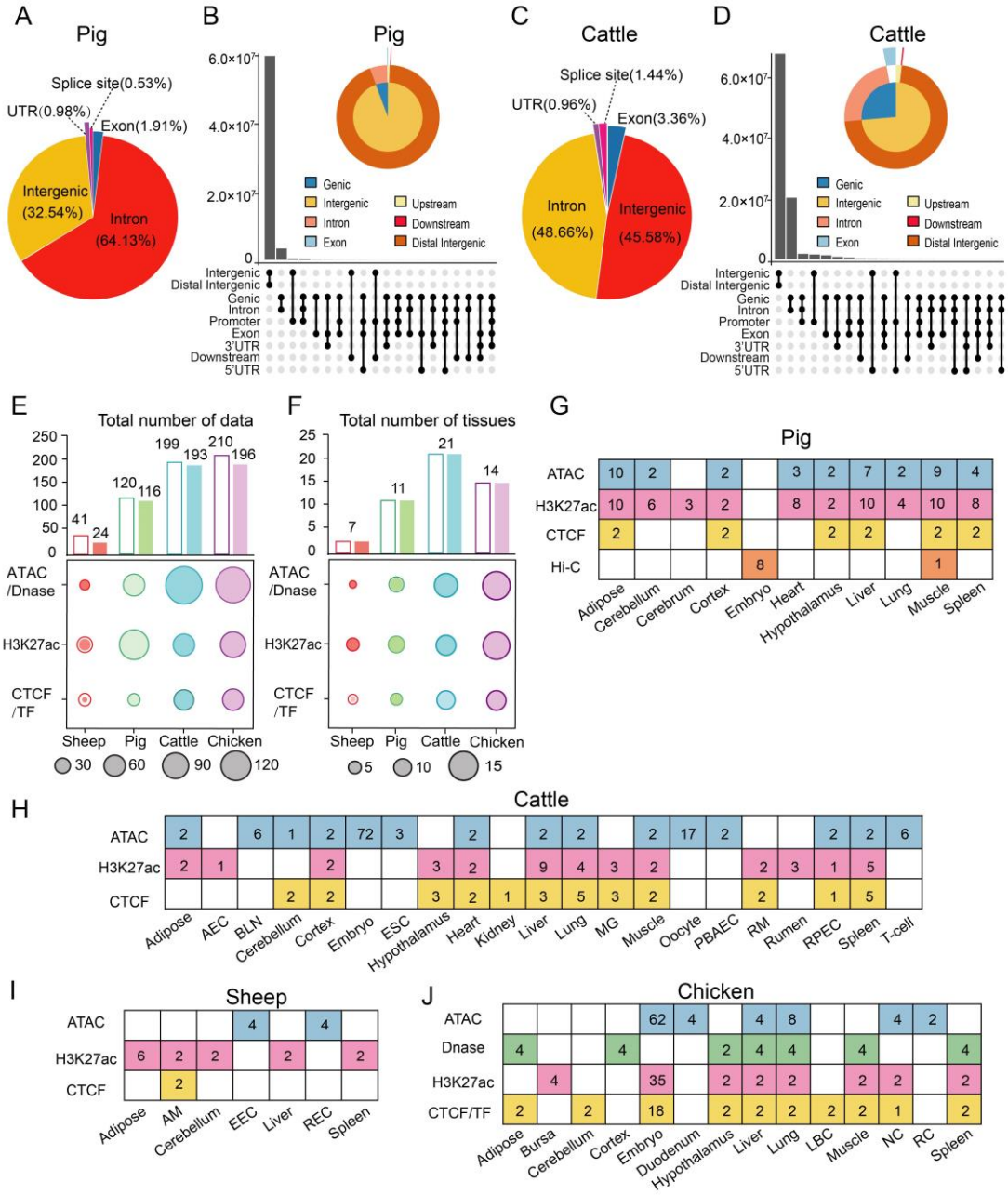

Figure 1. Genomic distribution of variants and their predicted effects on annotated genes and epigenomic datasets collection in livestock. *(A)* Percent distribution of variants in different regions in the susScr11 reference genome. *(B)* Regions within genome predicted to be affected by variants based on annotated genes in the susScr11; black dots indicate regions within a gene predicted to be simultaneously affected a variant. *(C)* Distribution of variants in the bosTau9 reference genome. *(D)* Predicted effects of cattle variants based on annotated genes in the bosTau9 genome assembly. *(E)* Statistical summary of epigenomic datasets for the four species. Histogram of total numbers of datasets obtained for each species; empty columns (left) are raw data and filled columns (right) show number of datasets after filtering and quality control. Bubble size represents number of different epigenetics datasets; outer circles, raw data; inner circles, cleaned

data. *(F)* Statistical summary of epigenetics datasets for different tissue types in the four species. Empty histograms (left) are number of tissue types with raw datasets; filled histograms (right) are number of tissue types with cleaned datasets. Bubble size indicates number of different tissues represented in each data type; outer circles are raw datasets; inner circles are cleaned datasets. *(G-J)* Summary of different quality-controlled epigenomic datasets and represented tissue types in *(G)* pig, *(H)* cattle, *(I)* sheep, and *(J)* chicken. AEC, aortic endothelial cells; BLN, bronchial lymph node; ESC, embryonic stem cells; MG, mammary gland; PBAEC, primary bovine aortic endothelial cells; RM, renal medulla; and RPEC, rumen primary epithelial cell. AM, alveolar macrophage; EEC, esophagus epithelium cells; and REC, rumen epithelium cells. LBC, lymphoma B-cell; NC, neural crest; RC, retinal cell.

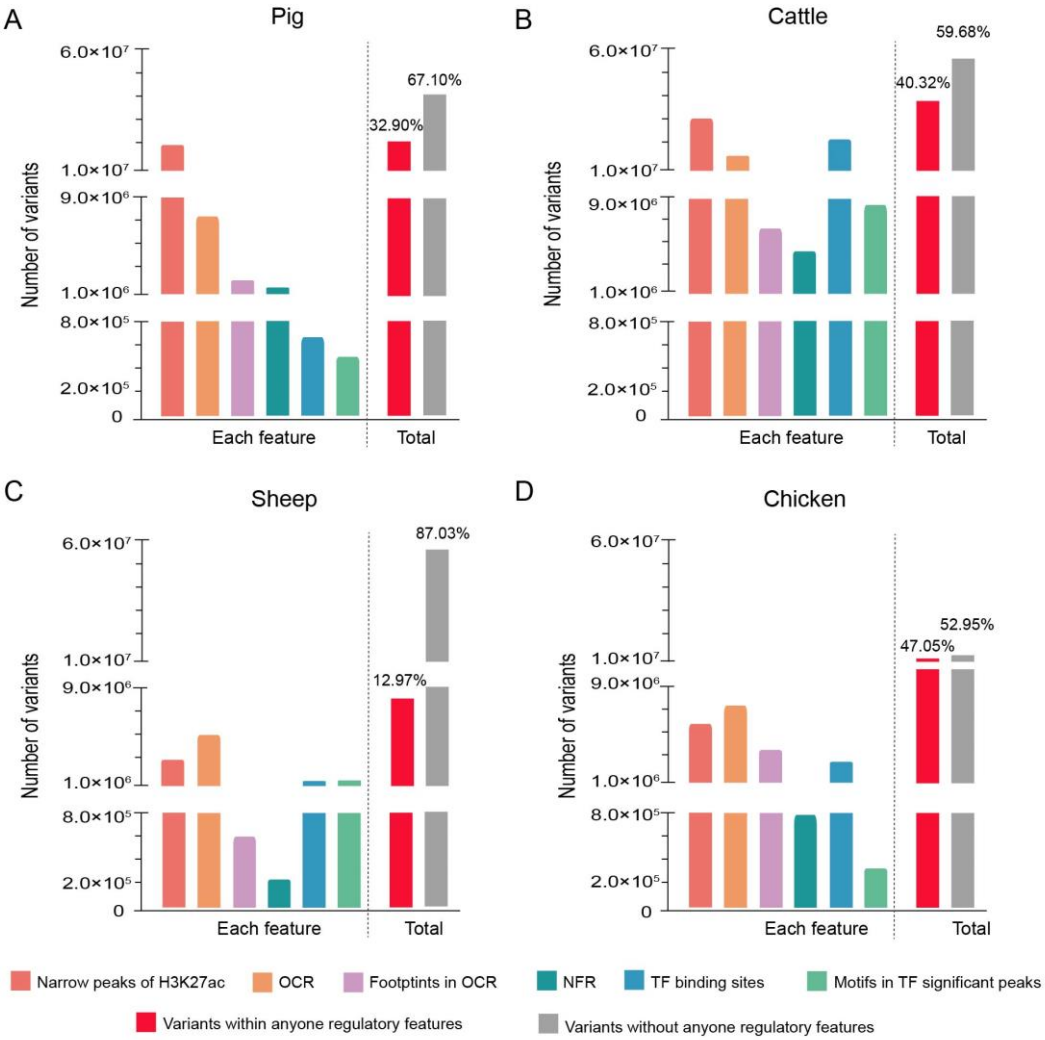

Figure 2. Statistical summary of candidate functional variants and their distribution in genomic regulatory features in four livestock species. *(A)* Total number of candidate functional variants distributed in each of 7 regulatory features, including narrow peaks in H3K27ac, open chromatin region (OCR), footprints in OCR, nucleosome free regions (NFR), recognition motifs in significant transcription factor (TF) peaks in ChIPseq, and TF binding sites, or associated with no regulatory features in epigenomic data from *(A)* pig, *(B)* cattle, *(C)* sheep, and *(D)* chicken.

Figure 3 revised

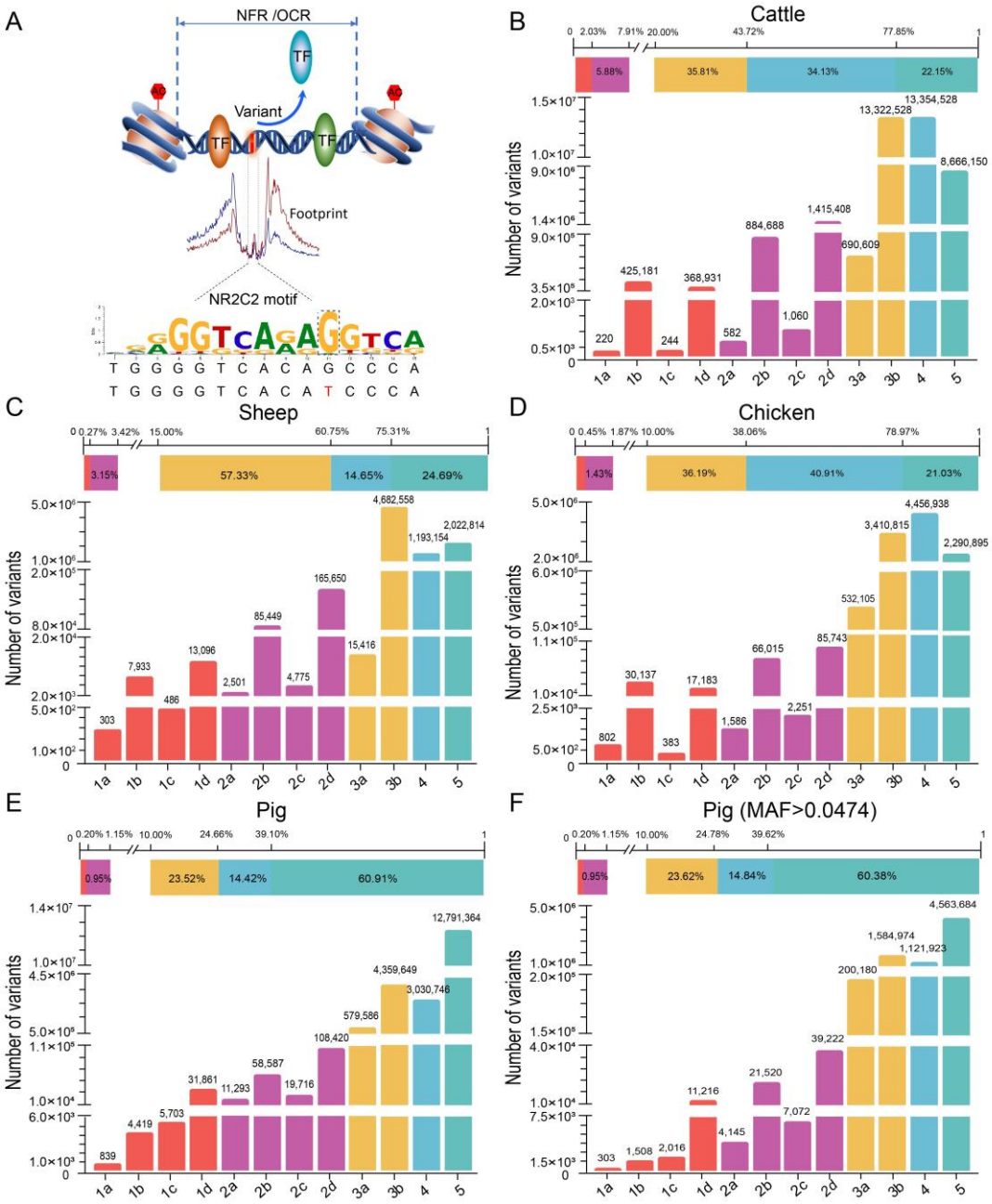

Figure 3. Confidence scoring of candidate functional variants. *(A)* Illustration of design principle of Functional Confidence scoring system. *(B-E)* Statistical summary of candidate functional variant distribution among confidence subcategories in *(B)* cattle, *(C)* sheep, *(D)* chickens, and *(E)* pigs. Bar at the top shows the proportional distribution of main confidence categories among total candidate functional variants for each livestock species. *(F)* Number of candidate functional variants in each subcategory filtered by minor allele frequency (MAF>0.047) from 491 whole genome sequencing datasets in pigs.

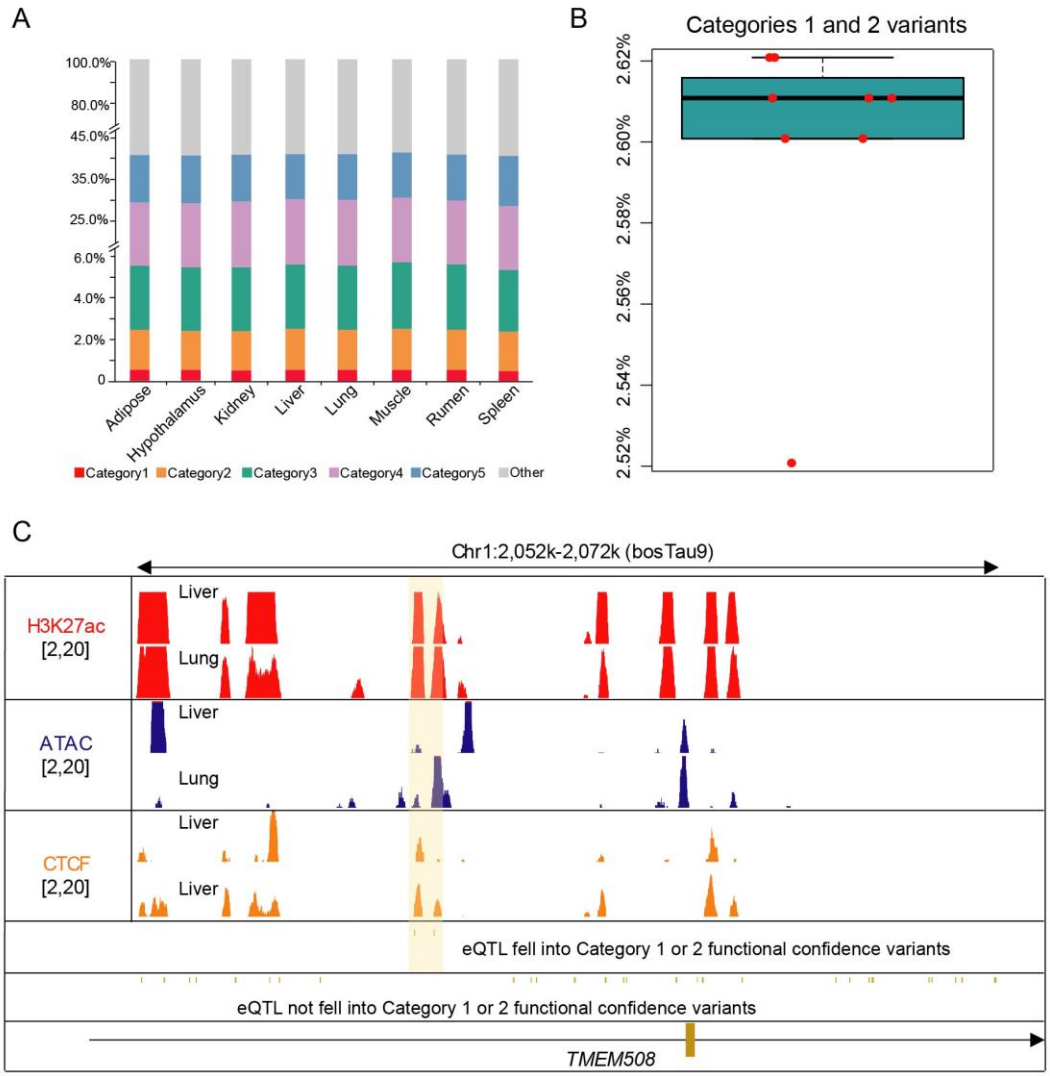

Figure 4. Assessment of cis-eQTL data from cattle using IFmut. *(A)* Distribution of cattle cis-eQTLs from 8 tissues (adipose, hypothalamus, kidney, liver, lung, muscle, rumen and spleen) in different Functional confidence categories assigned by IFmut. *(B)* Proportion of cis-eQTLs classified as high or moderate confidence candidate variants (Categories 1 and 2) in all cis-eQTLs. *(C)* Example of visualizing high and moderate confidence cis-eQTLs around the *TMEM508* gene in epigenomic data from cattle in JBorwse tool.

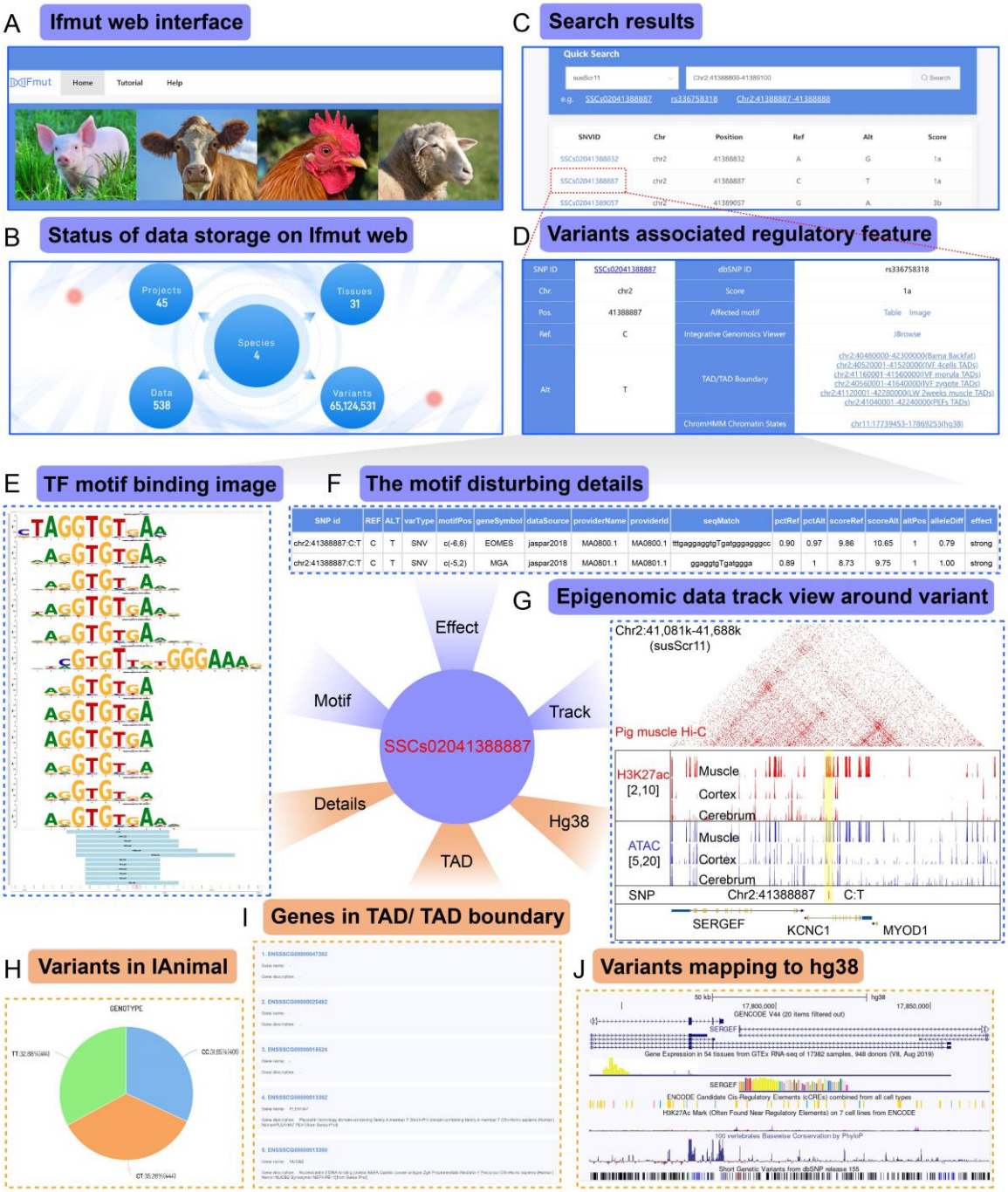

Figure 5. Overview of the Integrated Functional Mutation (IFmut) database. *(A)* IFmut home page, containing information about the four livestock species. *(B)* Graphic summary of datasets available in IFmut organized by search type. *(C)* Example of search results generated with the "Quick Search" function, including variant chromosomal location, conversion type, and confidence score. *(D)* Variant-associated regulatory features in the SNVID query results of Quick Search are linked to pages containing information such as motif affecting, TAD/boundary, and ChromHMM of human genome conservation region, which are further linked to source data, external databases, etc. *(E)* Clicking the "Image" link in the "Affected motif" column in *(D)* takes the user to logos plots of nucleotide conservation in TF recognition motif(s) potentially affected by a queried variant. *(F)* Clicking "Table" in the "Affected motif" column in *(D)* takes the user to a page containing the predicted effect on TF binding, and the affected gene symbol of the TF

motif etc. details about the potentially affected TF motif(s). *(G)* The track view of variant through "JBrowse" function in *(D)*. This function will bring the user to track views and feature visualization for regions containing the queried variants in ATAC-seq and ChIP-seq (H3K37ac) data, Hi-C interaction heatmaps (for pig), and nearby genes. *(H)* Clicking on SNVID hyperlinks in *(D)* brings the user to the IAnimal database (https://ianimal.pro/) to obtain additional information, such as genotype and major allele frequency in pig or cattle. *(I)* A subsequent search for genes within topologically associating domains (TADs) or TAD boundaries that contain user queried variant provides links to information about those genes in the IAnimal database (https://ianimal.pro/). *(J)* Users can also perform comparative genomics between predicted variant-affected regions in livestock and corresponding chromatin regions in the human hg38 reference genome mapped using LiftOver.

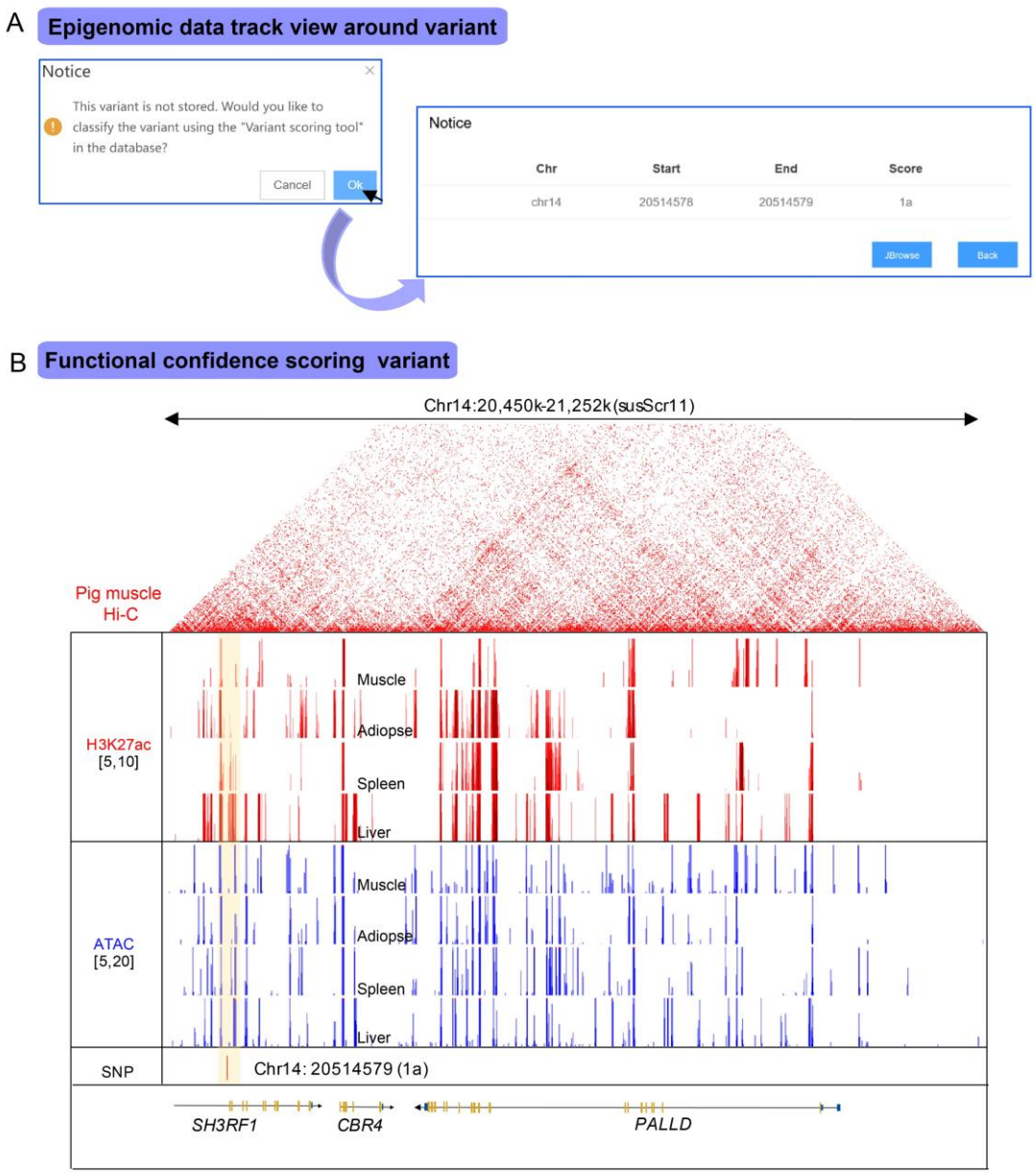

Figure 6. Functional confidence scoring and epigenomic data visualization functions in IFmut for user analysis of novel candidate variants. *(A)* Functional Confidence Scoring tool in IFmut. For variants of interest not stored in IFmut, users are prompted with the option to conduct Functional Confidence scoring using the tool in IFmut. *(B)* Epigenomic data visualization to assess novel variants. To examine the evidence underlying the IFmut functional confidence score for a variant of interest, users can follow a link to the JBrowse tool showing epigenomic tracks, TAD regions, and nearby genes.

Table 1. Number of regulatory features detected in livestock species.

Regulatory feature	pig	cattle	sheep	chicken
Basic regulatory feature				
NFR	200,450	328,909	58,655	103,504
OCR	441,818	780,224	418,560	600,364
Narrow H3K27ac peaks	163,307	162,841	57,167	69,527
Footprints in OCR	2,318,229	4,002,427	978,982	3,157,063
Total non-redundant	352,763	804,593	474,751	1,026,316
TF binding site-related features				
Footprint-matching motifs	125,500	1,352,480	253,476	1,098,796
Partial motif-containing footprints	198,272	1,569,613	374,664	680,374
Motifs in significant peaks	1,354,404	24,540,298	15,242,424	2,962,352
TF binding sites	43,007	182,643	136,538	135,362
Total non-redundant	451,212	2,783,299	2,443,102	428,545

NFR, nucleosome-free regions; OCR, open chromatin regions.

Table 2. Genomic coverage of regulatory features.

Regulatory feature	pig	cattle	sheep	chicken
Basic regulatory feature (bp)				
NFR	60,641,955	95,978,174	9,085,698	33,571,036
OCR	300,233,553	324,314,161	186,787,025	309,414,106
Narrow H3K27ac peaks	700,812,854	608,126,611	117,598,663	229,512,492
Footprints in OCR	93,287,631	127,134,565	24,974,332	151,709,670
Total non-redundant	787,386,263	794,764,318	287,920,733	441,600,747
TF binding site-related features (bp)				
Footprint-matching motifs	3,295,980	22,900,815	4,146,394	18,530,843
Partial motif-containing footprints	7,771,203	46,493,819	9,943,498	30,423,419
Motifs in significant peaks	20,463,073	135,948,968	53,057,694	50,400,477
TF binding sites	31,158,919	332,861,617	55,897,125	100,867,511
Total non-redundant	51,963,305	391,146,808	104,501,330	107,662,441

NFR, nucleosome-free regions; OCR, open chromatin regions.

Table 3. Variant classification scheme for scoring system in IFmut database.

Categorization Scheme	
Category	**Description**
	High possibility of affecting transcription factor binding
1a	QTL + OCR + NFR + Footprints in OCR + Footprint-matching motifs
1b	OCR + NFR + Footprints in OCR + Footprint-matching motifs
1c	QTL + OCR + NFR + Part motif-containing footprints
1d	OCR + NFR + Part motif-containing footprints
	Moderate possibility of affecting TF binding
2a	QTL + OCR + Footprints in OCR + Footprint-matching motifs
2b	OCR + footprint in OCR+ Footprint-matching motifs
2c	QTL + OCR + Part motif-containing footprints
2d	OCR + Part motif-containing footprints
	Low possibility of affecting TF binding
3a	OCR + NFR + Footprints in OCR
3b	OCR / NFR + Motifs in significant peaks
	Minimal possibility of affecting TF binding
4	OCR / NFR / Footprints in OCR / TF binding significant peaks
	Likely to be associated with gene expression
5	H3K27ac significant peaks

QTL, quantitative trait loci; NFR, nucleosome-free regions; OCR, open chromatin regions

Table 4. Predictive reliability of EBV for traits ADG and BF.

Scenarios	SNP numbers	ADG	BF
11k random	11,000	0.268 (0.014)	0.265 (0.015)
muscle	10,544	0.319	0.316
liver	6,049	0.348	0.346
adipose	3,801	0.380	0.378

Numbers within the parentheses are the standard errors of the reliability.