UNIVERSITE DE LIEGE

INSTITUT GIGA & FACULTE DE MEDECINE VETERINAIRE

UNITE DE GENOMIQUE ANIMALE

Contribution de différentes classes fonctionnelles à la variation génétique et leur utilisation pour améliorer la fiabilité des prédictions génomiques

Contribution of different functional classes to genetic variation and their use to improve reliability of genomic predictions

Can YUAN

THESE PRESENTEE EN VUE DE L'OBTENTION DU GRADE DE DOCTORAT EN SCIENCES VETERINAIRES

ANNEE ACADEMIQUE 2023-2024

Acknowledgement

"Swift fly the years; my four-year PhD journey has come to an end—a challenging, wonderful, and rewarding experience. I would like to express my gratitude to those whose support was indispensable for the completion of this thesis. In particular, I owe a special thanks to Tom and Michel. They generously spent plenty of time to guide me in all the three studies. Tom dedicated an enormous amount of time to meticulously read and revise the thesis."

Firstly, I must express my sincere gratitude to my supervisory team. Michel and Haruko, thank you for your invaluable guidance in the establishment of ATAC-Seq resources which are the basis for present study. Tom and Carole, thank you for your supervising and unwavering support since the first day you welcomed me as PhD student in your group. You were active in conceiving this study, guiding my experiments and revising the manuscript. Your expertise and patient assistance made all the difference. The completion of this comprehensive project that involve knowledge in functional genomics, bioinformatics and quantitative genetics would not have been possible without your generous support. I benefit a lot in and beyond science with such an excellent advisory team. Your wisdom, rigorousness and patience not only had a vital role to overcome challenges I encountered but also contributed tremendously to my growth and improved my ability towards my future scientific career. These are the memories I never want to forget: you educated me how to write, you inspired me to consistently keep learning and thinking, and you trained me to present my result clearly.

I would like to acknowledge my thesis jury and its president, Matthew Robinson, Andrea Rau, Nicolas Gengler, Pierre Geurts, Luc Grobet, Anne-Sophie Van Laere and Benjamin Dewals, for accepting to evaluate my thesis. I would also like to thank Frédéric Farnir from the thesis committee for the annual evaluation, which guaranteed that the research progress remained on track.

The Walloon Breeders Association deserves special thanks for the stimulating collaboration, for sharing their phenotypic and genotypic and sequence data as well as for their financial support.

I would like to extend my gratitude to all colleagues from GIGA Medical genomics, in particular, the members of the Unit of Animal Genomics: Jose-Luis, Natalia, Maulana, Claire, Agnieszka, Caroline, Gabriel, Lim, Lijing, Yefang, Wouter, Latifa. These brilliant friends and colleagues who improve my coding, share your expertise and give me endless courage and strength to conquering difficulties. Our debate in seminar, chat in cafeteria and drink time in pub, I will always cherish these moments and the love of all of you in my heart. I am especially grateful to Miyako for your kind help that made my adventure in Liege much easier. There are no proper words to convey my incredibly gratitude for all your help and concern. I feel so lucky to have been a member of UAG.

I am truly thankful for all the collaborators for their tremendous efforts in sample collection, technical support and fruitful discussion. I am grateful for the support of the GIGA Genomics and Bioinformatics team and the CÉCI cluster team.

My sincere thanks would go to all my family and friends in China and Belgium.

Abbreviations

AI	Average Information
ATAC-Seq	The assay for transposase-accessible chromatin with sequencing
BBB	Belgian Blue Beef
BBB	Belgian Blue Beef
BBC	Belgian Blue Cattle
BMI	Body mass index
BSLMM	Bayesian Sparse Linear Mixed Model
CAD	Cardiovascular disease
CDS	Coding sequence
cGTEx	cattle GTEx
ChIP-seq	ChIP sequencing
ChromHMM	Hidden Markov model
CLCN7	CIC-7 Chloride Channel
CMD1	Congenital muscular dystonia 1
CMD2	Congenital muscular dystonia 2
CNS	Central nervous system
CP	Collective peak
CpG	Cytosine nucleotides followed by guanine nucleotides
CTS	Crooked tail syndrome
CVs	Causal variants
DBDs	DNA-binding domains
DHS	DNaseI hypersensitivity sites
DNase-seq	DNase I hypersensitive sites sequencing
DNMs	De novo mutations
EM	Expectation Maximization
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression quantitative trait loci
EST	Expressed sequence tag
FAANG	Functional Annotation of Animal Genomes
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements followed by sequencing
FAN1	First functional annotation model
FAN2	Second functional annotation model
FarmGTEx	Farm animal Genotype-Tissue Expression
FRip	fraction of reads in called peak regions
FUN1	First functional marker panel
FUN2	Second functional marker panel
FUN3	Third functional marker panel
GBLUP	Genomic best linear unbiased prediction
GBLUP-C	GBLUP based on centered genotype
GBLUP-S	GBLUP based on standardized genotype
gDNA	Genomic DNA
GEBV	Genomic estimated breeding values
GENE-SWitCH	GENomE of SWine and Chicken

GERP	Genomic evolutionary rate profiling
GF	Genomic features
GFBLUP	Genomic feature BLUP
GO	Gene ontology
GREML	Genomic restricted maximum likelihood
GRM	Genomic relationship matrix
GS	Genomic selection
GTEx	Genotype-Tissue Expression
GTF	General Transfer Format
GWAS	Genome-wide association studies
H3K27ac	histone H3 lysine 27 acetylation
H3K4me3	histone H3 lysine 4 methylation
H3K9me3	histone H3 lysine 9 trimethylation
HMD	High Marker Density
ICM	Inner cell mass
IDR	Irreproducible discovery rates
IHEC	International Human Epigenome Consortium
indels	Insertion and deletions
IP	Individual peak
LALBA	Lactalbumin alpha
LD	Linkage disequilibrium
LDMS	LD and MAF-stratified
LDRS	LD regression score
LDS	LD-stratified
LDSC	LD score regression
LMD	Low Marker Density
LMM	linear mixed models
lncRNAs	long non-coding RNAs
MAF	Minor allele frequency
MAS	Marker assisted selection
Mbp	Megabase Pair
MC	Multiple component
MGFBLUP	Multiple genomic feature BLUP
MGFBLUP-C	Multiple genomic feature BLUP based on centered genotype
MGFBLUP-S	Multiple genomic feature BLUP based on standardized genotype
miRNAs	MicroRNAs
MMD	Medium Marker Density
MNase-seq	Micrococcal nuclease sequencing
modENCODE	Model Organism ENCyclopedia Of DNA Elements
MRC2	Mannose receptor C type 2
MS	MAF-stratified
MSTN	Myostatin
MYH1	Myosin heavy chain 1
NMF	non-negative matrix factorization
noLDMS	Without correction for MAF or LD score
OCR	Open chromatin regions

ORFs	Open reading frames
PIC	Transcription preinitiation complex
PIGH	phosphatidylinositol glycan anchor biosynthesis class H
PIP	Posterior inclusion probabilities
PWMs	Position weight matrices
QTL	Quantitative trait locus
QTLs	Quantitative trait loci
QTN	Quantitative trait nucleotide
REML	Restricted maximum likelihood
RMSE	Residual mean square error
RNF11	Ring Finger Protein 11
scATAC-Seq	Single cell ATAC-Seq
SFS	Site frequency spectrum
SNP	Single nucleotide polymorphism
SNPs	Single nucleotide polymorphisms
SNVs	Single nucleotide variants
sQTLs	splicing QTLs
ssGBLUP	single step GBLUP
T2D	Type 2 diabetes
TC	Two component
TF	Transcription factor
TFBSs	Transcription factor binding sites
TSS	Transcription start sites
TTS	Transcription termination sites
UDR	Upstream or downstream of genes
US	Unstratified
UTR	Untranslated regions
VCF	Variant Calling file
VEP	Variant Effect Predictor
vWF	von Willebrand factor
WGS	Whole genome sequence
ZF	Zinc finger

Abstract - Résumé

Résumé

Depuis qu'elle a été proposée en 2001, la sélection génomique (SG) a été mise en œuvre avec succès dans plusieurs grandes espèces d'élevage. Cependant, ses performances peuvent encore être améliorées, par exemple en termes de précision. Les modèles couramment utilisés donnent en effet le même poids à tous les variants, malgré leurs différences biologiques. Il a cependant été démontré que l'ampleur de l'effet d'un variant varie considérablement en fonction de sa catégorie fonctionnelle, les variants codants ayant des effets plus importants. En outre, des études récentes sur les caractères complexes chez l'homme ont mis en évidence l'importance des variants régulateurs qui affectent l'expression des gènes. De même, l'inclusion d'informations d'annotation fonctionnelle dans les modèles de prédiction génomique a permis d'améliorer leur précision chez l'homme. Enfin, l'identification des éléments régulateurs et des variants régulateurs est également essentielle pour l'identification des variants causaux dans les études d'association.

Les objectifs de cette thèse étaient de contribuer à l'annotation fonctionnelle du génome bovin en générant un catalogue de variants régulateurs, et d'utiliser cette annotation fonctionnelle pour étudier l'importance de différentes catégories fonctionnelles dans la variation génétique de caractères complexes d'intérêt chez les bovins blanc bleus belges (BBB). Enfin, j'ai cherché à savoir si l'incorporation de ces informations fonctionnelles dans les modèles de prédiction génomique pouvait améliorer leur précision.

Dans la première partie de cette étude, 104 échantillons provenant de 63 types de tissus ont été analysés par ATAC-Seq, l'une des approches les plus utilisées pour la détection d'éléments régulateurs. Au total, 976 813 pics ont été détectés, représentant 10,0 % du génome. Les pics proximaux (c'est-à-dire proches des gènes) étaient plus ouverts et actifs dans un plus grand nombre de types de tissus que les pics distaux, qui sont plus éloignés des gènes et plus spécifiques aux tissus. Les scores GERP, qui reflètent les contraintes évolutives, ont une distribution dispersée dans les régions de chromatine ouverte (RCO), ce qui suggère un processus évolutif complexe. Pour faciliter l'interprétation biologique des pics observés, nous avons compressé la matrice contenant les 976 813 pics de 104 échantillons en 16 composantes par factorisation matricielle non négative. La plupart des composantes résultantes ont pu être facilement attribuées à un processus biologique car elles regroupaient des tissus anatomiquement et fonctionnellement similaires. Nous avons ensuite caractérisé les variants dans les RCO en utilisant les données de séquence de 264 individus Holstein. Sur les 11030905 variants nucléotidiques détectés dans cette cohorte, 1256997 correspondaient à des pics ATAC-Seq. Malgré le nombre plus faible de variants attendus dans les RCO en raison de la sélection purifiante, nous avons observé un niveau plus élevé de polymorphismes. Cela pourrait être dû à un taux de mutation plus élevé dans les RCO, hypothèse supportée par la fréquence des 'singletons' et des mutations de novo 1,3 fois plus élevée dans ces régions. Nous avons ensuite utilisé 7 817 eQTLs sanguins et 6 172 eQTLs hépatiques pour évaluer si les RCO étaient enrichies en variants régulateurs. En utilisant une méthode de permutations pour quantifier l'enrichissement, nous avons constaté que les eQTLs ont tendance à chevaucher les RCO plus

souvent que par hasard, et que cet enrichissement est spécifique au tissu. Nous avons estimé que la proportion de variants régulateurs correspondant aux pics ATAC-Seq est d'environ 1 sur 3, et que la proportion de variants dans les pics ATAC-Seq qui sont régulateurs est d'environ 1 sur 25, ce qui suggère que ce catalogue peut être utile pour la SG.

L'objectif de la deuxième partie de la thèse était d'estimer la proportion de variance génétique associée aux éléments régulateurs pour les caractères de développement musculaire en utilisant environ 15000 vaches BBB dont les génotypes ont été imputés au niveau de la séquence. Bien que des méthodes telles que GREML et BayesRR-RC permettent un tel partitionnement de l'héritabilité, leur précision n'a pas été étudiée de manière approfondie chez les espèces animales, qui ont une structure très différente de celle des données humaines en termes de taille effective de la population, de niveaux de parenté et de déséquilibre de liaison (DL). Nous avons donc évalué ces méthodes sur la base d'une étude de simulation utilisant la structure de nos données. En l'absence de stratification, nous avons constaté que les méthodes étaient non biaisées et imprécises. Comme chez l'homme, des modèles plus complexes avec davantage de catégories de variants étaient nécessaires en présence d'une stratification due au DL ou aux fréquences alléliques. Dans des scénarios plus complexes, avec plusieurs groupes d'annotation ayant des tailles d'effet différentes, nous avons observé qu'il pouvait y avoir une confusion entre les différentes catégories et que les estimateurs étaient imprécis. Néanmoins, ils étaient toujours indicatifs des tendances globales. Enfin, lorsque les méthodes ont été appliquées à des caractères de développement musculaire, nous avons constaté que les RCO contribuaient significativement à la variance génétique et que les variants codants avaient les effets les plus importants par SNP, ce qui plaide en faveur de l'utilisation de ces informations dans la GS.

Dans le dernier chapitre, j'ai évalué les avantages de l'utilisation des données de séquence et d'annotation fonctionnelle pour améliorer la précision de la SG. L'utilisation de la séquence a légèrement augmenté la fiabilité de la SG, et une légère augmentation supplémentaire a été observée avec lors de l'incorporation de l'annotation fonctionnelle dans le modèle GBLUP. J'ai ensuite testé une stratégie alternative en sélectionnant des sous-ensembles de marqueurs. Cette stratégie a permis d'utiliser davantage de modèles. En particulier, le Bayesian Sparse Linear Mixed Model (BSLMM), combinant un effet polygénique avec quelques variants majeurs, a permis d'obtenir une plus grande précision pour tous les caractères. Les meilleurs résultats ont été obtenus lorsque les informations fonctionnelles ont été utilisées pour sélectionner les marqueurs. J'ai également observé qu'une plus grande précision était obtenue avec des génotypes centrés qu'avec des génotypes standardisés, ce qui indique une relation différente entre la taille de l'effet du marqueur et la fréquence de l'allèle que chez l'homme, probablement en raison de la sélection directionnelle et de la présence de quelques variants communs à grand effet.

Pour conclure, cette thèse a démontré l'importance des variants régulateurs chez les bovins et qu'ils peuvent être utilisés pour améliorer la précision de la SG, bien qu'il reste des progrès à faire.

Summary

Since it was first proposed in 2001, genomic selection (GS) has been successfully implemented in several major livestock species and has had a major impact on livestock breeding. However, its performance can be further improved, for example in terms of accuracy. The models commonly used in GS give equal weight to all variants, despite their biological differences. In fact, the effect size of a variant has been shown to vary significantly as a function of the functional category of the variant, with coding variants having larger effects. In addition, recent studies of complex traits in humans have highlighted the importance of regulatory variants that affect traits by perturbing gene expression in a quantitative manner. Consistently, the inclusion of functional annotation information in genomic prediction models has led to improved prediction accuracy in humans. Finally, the identification of regulatory elements and regulatory variants is also essential for the identification of causative variants in association studies.

The objectives of this thesis were to contribute to the functional annotation of the bovine genome by generating a catalog of regulatory variants, and to use this functional annotation to study the importance of different functional categories in the genetic variation of complex traits of interest in Belgian Blue cattle (BBC). Finally, I investigated whether incorporating this functional information into genomic prediction models could improve their accuracy when applied to the same set of traits.

In the first part of this study, 104 samples from 63 tissue types were subjected to transposase accessible chromatin using sequencing (ATAC-Seq), one of the most widely used approaches for the detection of regulatory elements. A total of 976,813 peaks were detected, representing 10.0% of the genome (5% for the core regions of the peaks). Proximal peaks (i.e., close to genes) were more open and active in more tissue types than distal peaks, which are more distant from genes and more tissue specific. Genomic evolutionary rate profiling (GERP) scores, reflecting evolutionary constraints, had a dispersed distribution in open chromatin regions (OCR), suggesting a complex evolutionary process. To facilitate biological interpretation of the observed peaks, we compressed the matrix containing the 976,813 peaks from 104 samples into 16 components by non-negative matrix factorization. Most of the resulting components could be easily assigned to a biological process as they grouped anatomically and functionally similar tissues. We then characterized variants in OCR using sequence data from 264 Holstein individuals. Of the 11,030,905 single nucleotide variants detected in this cohort, 1,256,997 mapped to ATAC-Seq peaks, including 847,831 common variants. Despite the lower expected number of variants in OCR due to evolutionary constraints resulting from purifying selection, we observed a higher level of polymorphisms. This could be due to an increased mutation rate in OCRs, a hypothesis supported by a 1.3-fold increase in the frequency of singletons and de novo mutations in these regions, in agreement with recent findings in humans and Arabidopsis. We then used 7,817 blood and 6,172 liver expression quantitative trait loci (eQTLs) and their credible sets, obtained from an experiment with more than 170 samples, to evaluate whether OCR were enriched in the regulatory variant. Using a

permutation-based method to quantify enrichment, we found that credible sets tend to overlap with ATAC-Seq more often than by chance alone, and that this enrichment is tissue-specific. We estimated that the proportion of regulatory variants mapping to ATAC-Seq peaks is approximately 1 in 3, and that the proportion of variants in ATAC-Seq peaks that are regulatory is approximately 1 in 25, suggesting that this catalog may be useful for GS.

The objective of the second part of the thesis was to estimate the proportion of genetic variance associated with regulatory elements for muscular development traits in BBC using a cohort of ~15,000 BBC cows with imputed genotypes at the sequence level. Although methods such as genomic restricted maximum likelihood (GREML) and BayesRR-RC allow for such heritability partitioning, their accuracy has not been thoroughly studied in livestock species, which have a very different structure compared to human data in terms of effective population size, relatedness levels, and linkage disequilibrium (LD) patterns. Therefore, we started an evaluation of these methods based on a simulation study using the structure of our data sets. Overall, we found that the methods were unbiased and imprecise in the absence of stratification. As in humans, more complex models with more categories of variants were required in the presence of LD or allele frequency stratification. In more complex simulation scenarios, with multiple annotation groups having different effect sizes, we observed that there could be confounding between the different categories and that the estimators were imprecise. Nevertheless, they were still indicative of global trends. Finally, when the methods were applied to muscular development traits, we found that OCR accounted for a large fraction of the genetic variance and that coding variants had the largest effects per single nucleotide polymorphisms (SNPs), arguing for the use of this information in GS.

In the final chapter, I evaluated the benefit of using whole-genome sequence (WGS) data and functional annotation to improve the accuracy of GS, with a particular emphasis on coding and regulatory variants and on elements detected in muscle, the tissue of interest in BBC. Using WGS data slightly increased the reliability of GS, and a slight additional increase was observed with GBLUP when incorporating functional annotation into the model, but not with BayesRR-RC. I then tested an alternative strategy by selecting subsets of markers. This strategy allowed more models to be used, as GBLUP and BayesRR-RC were the only approaches that could be run on the full sequence. In particular, the Bayesian Sparse Linear Mixed Model (BSLMM), a model that fits a polygenic effect combined with a few large effect variants, achieved the highest reliabilities across all traits. The best results were obtained when the functional information was used to select the marker panels. I also observed that higher accuracy was obtained with centered genotypes compared to standardized genotypes, indicating a different relationship between marker effect size and allele frequency than in humans, probably due to directional selection and the presence of a few common large-effect variants. Here, only 50 to 200 large-effect variants were fitted by BSLMM in the best models.

Overall, the thesis has shown the importance of regulatory variants in cattle and that they can be used to improve the accuracy of GS although there is still room for improvement.

General preamble

Animal breeding has seen significant advances over the years, from pedigree based breeding value estimation, through marker assisted selection (MAS), to genomic selection, enabling continuous increases in livestock productivity (Georges et al., 2019; Meuwissen et al., 2016). In recent years, genomic selection has been widely adopted and has had a tremendous impact on livestock breeding, including changes in breeding programs (e.g. García-Ruiz et al., 2016). The key feature of genomic selection, proposed by Meuwissen et al. in 2001, is its ability to use dense variant maps that allow the effects of quantitative trait loci (QTLs) associated with the trait of interest to be captured. Its main advantages are increased accuracy of breeding values coupled with a reduced generation interval, and the ability to select for traits that are more difficult to measure or for sex-specific traits (Goddard and Hayes, 2007; Meuwissen et al., 2016). The cost of genotyping and sequencing technologies has fallen dramatically over the last 20 years, making it possible to genotype large cohorts of individuals and perform large-scale resequencing of entire genomes on hundreds of individuals (Quackenbush, 2022). Accordingly, the number of animals genotyped and whole genome sequenced has steadily increased over the last decade, greatly facilitating the use of genomic selection. Combining these data with efficient missing genotype imputation techniques allows genome-wide association studies and genomic selection to be performed at the sequence level in large cohorts, with the potential to increase the power of association studies and the accuracy of genomic predictions (Quick et al., 2020). In cattle, for example, the whole genome sequence data from 2,703 individuals covering the major ancestors of several breeds gathered in the 1000 Bull Genomes Project provides a valuable resource for genotype imputation, greatly accelerating the identification of causative variants (Daetwyler et al., 2014).

Nevertheless, the accuracy of genomic selection using sequence-level data remains inferior to what can be achieved given the heritability of the selected traits. This could have a number of causes, including the size and composition of the reference population, the accuracy of imputation or the contribution of dominance and epistasis to the genetic architecture of the traits of interest. Another factor is that all variants are generally given equal weight in the computation of the additive relationship between animals required for GBLUP analyses, or equal prior probabilities of variant effects in Bayesian approaches. Yet, only a minority of variants are causative, the remainder being at best passenger variants in LD with one or more of the causative variants. The extent of LD between causative and passenger variants is bound to be population or even sub-population specific and is likely to vary over time, which may partly account for the observed limitations in selection accuracy. It is generally accepted that causative variants are essentially coding and regulatory variants, although our understanding of complex traits remains partial and other mechanisms may be at play. Coding variants, including missense, nonsense, frameshift, splice site variants and deletions, are easily identified using bioinformatic annotation tools. They account for only a small proportion of the genome, and their contribution to genetic variance of complex phenotypes in livestock species, has been estimated to be modest (Koufariotis et al., 2014). Regulatory variants can act either by perturbing the expression profile of genes located in cis, or possibly, by perturbing the gene regulatory network and affecting the expression profile

of a restricted number of core genes in trans. Therefore, comprehensive genome annotations, which allows the identification of coding and more importantly regulatory variants, are critical for identifying the causative variants. More generally, it is widely accepted that variants in different functional classes contribute heterogeneously to genetic variation (Finucane et al., 2015). To gain further insight into the genetic architecture of complex traits in humans and livestock, heritability partitioning methods have been extensively used in combination with available functional annotations, although further investigation is needed to estimate the exact contribution of different functional categories to genetic variance, especially in livestock. Such knowledge would facilitate the use of biological priors to prioritize or weight variants to further improve the accuracy of genomic prediction.

Several collaborative efforts have been undertaken to identify functional elements and variants in the human genome, including the Encyclopedia of DNA Elements (ENCODE) (de Souza, 2012), Roadmap Epigenomics (Kundaje et al., 2015) and the Genotype-Tissue Expression (GTEx) (GTEx Consortium, 2020) projects. Results from these projects have highlighted the importance of regulatory elements, which have been found to underlie a large proportion of (Genome-wide association studies) GWAS hits and contribute substantially to the heritability of complex diseases and traits, confirming the potential benefit of prioritizing these variants in genome prediction. Consistently, more and more tools are being developed to incorporate functional annotations as priors in genomic prediction models. In animal species, similar consortia such as Functional Annotation of Animal Genomes (FAANG) (Andersson et al., 2015), Farm animal Genotype-Tissue Expression (FarmGTEx) (Liu et al., 2022), the regulatory GENomE of SWine and Chicken (GENE-SWitCH) (Acloque et al., 2022) and BovReg (Moreira et al., 2022) have produced similar catalogues of functional elements for livestock and poultry genomes. As in humans, heritability partitioning approaches using the available biological information revealed an enrichment of coding and regulatory variants among those associated with complex traits (Bhuiyan et al., 2018; Koufariotis et al., 2014). However, to date, functional annotations for regulatory elements in cattle are only available for a few tissues and developmental stages, which is insufficient for biological interpretation and severely limits their use in understanding the mechanism underlying complex traits.

The overall objective of my thesis is to better understand the contribution of different functional classes, with a focus on regulatory variants, to genetic variation and how this information can be used to improve reliability of genomic predictions. Therefore, in my thesis I have (1) contributed to the generation of a comprehensive catalogue of regulatory elements for the bovine genome to complement its functional annotation and tested its sensitivity and accuracy in identifying regulatory variants, (2) evaluated heritability partitioning methods in cattle and estimated the genetic contribution of different functional categories to the heritability of complex traits in Belgian Blue cattle (BBC), (3) evaluated strategies for using biological information, including the newly discovered regulatory elements, to improve genomic predictions. My manuscript begins with an introduction that gives an overview of these aspects in livestock and presents the BBC breed.

Introduction

1 Introduction

1.1 Belgian Blue beef cattle

1.1.1 The breed

Belgian Blue cattle (BBC), a breed found mainly in Belgium, are famous for their extreme muscle development, known as "double muscling" (see example in Figure 1.1), which results from a significantly increased number of muscle fibres rather than an increased fibre diameter (Grobet et al., 1997). The BBC originated from a cross between a local breed and a British Shorthorn breed, initially to improve the local breed for both dairy and beef production (Cheville, 1999; Kolkman, 2010). At the time, the population was small, with only about three thousand individuals reported. In the 20th century, market demand for meat products gradually shifted the focus of selection from dual-purpose to a more beef-oriented breed (Kolkman, 2010). At the same time, BBC individuals with the double-muscling phenotype, which is associated with increased muscle mass, became more common, highlighting their potential to improve meat production (Hanset, 2004; Kolkman, 2010). Intensive selection, combined with the widespread use of double-muscling bulls as sires through artificial insemination, accelerated the fixation of the trait and made it the iconic characteristic of the breed. Grobet et al. (1997) demonstrated that this double muscling trait was due to an 11-bp deletion in the myostatin gene (*MSTN*). However, there were problems with parturition in double-muscled cows and reduced semen quantity and quality in bulls (De Tavernier et al., 2001; Hoflack et al., 2008). Despite these challenges, the population increased significantly from 40,000 in 1970 to 450,000 in 1994, marking the first stage of genetic improvement as the use of this mutation was driven by the visible phenotypic changes in homozygous individuals. Research has shown that after 1985 this natural mutation is almost fixed in the new BBC population (Druet et al., 2014a). Interestingly, a significant variation in muscular traits has been observed in the new BBC population and further genetic gains have been achieved through genetic improvement (e.g. Druet et al., 2014a). Notably, the heritability of these traits remains above 30% even when excluding the effect of the double muscling mutation, which is now fixed.

Figure 1.1. A double-muscled Belgian Blue sire. The bull is Adajio de Bray, born in 2007. Figure downloaded from https://belgianbluegroup.com/bull/adajio-de-bray/

1.1.2 Genetic evaluation in Belgian Blue beef cattle

Until recently, genetic evaluation in BBC was based on pedigrees and progeny testing. Traits are mainly measured at three time points, including a first visit at birth, a second visit at around 14 months and a linear classification in adult cows (Hanset, 2004). At birth, conformation, gestation length and birth weight are recorded, while at the second visit, individuals are phenotyped for height, weight (predicted from thoracic perimeter) and conformation. In addition to these traits, so-called functional traits, such as drinking ability and defects, are reported. The heritability of these traits is not always high as the phenotypes can have some drawbacks. For example, until recently, birth weight has been evaluated by the eye of the breeder, whereas gestation length is influenced by caesarean section.

Linear scoring consist of visual inspection of the cows by trained technicians. Animals are scored (from 0 to 50) for 19 individuals traits that are grouped in three categories. In addition, skin thickness is manually assessed and height is measured. As a result, 7, 6 and 8 phenotypes are related to the size of the animal, the muscle development and to the bone structure and posture, respectively (see Figure 1.2). For genetic evaluation, multiple-trait models with canonical transformation are used (Hanset, 2004).

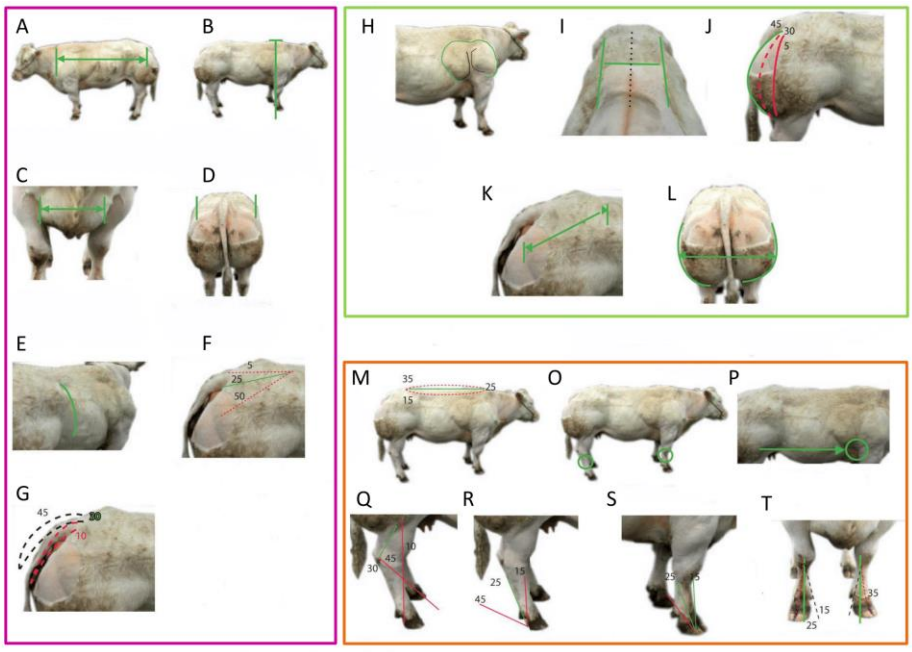


Figure 1.2. Linear classification traits recorded in adult cows. A-G. Traits related to body dimension including body length (A), height (B), chest width (C), pelvis width (D), rib shape (E), rump (F) and tail set (G). H-L. Traits related to muscular development including shoulder muscling (H), top muscling (I), buttock muscling – side view (J), pelvis length (K) and buttock muscling – rear view (L). Traits related to bone structure and posture including top line (M), bone (O), shoulder bone (P), hock (Q) and legs stance (R-T). Figure adapted from http://www.belgianblue.cz/fr/index.php?page=page&kid=22

The use of these breeding values has allowed a further increase in muscle mass, but this selection has also resulted in a negative trend in the height and length of the animals (Figure 1.3). The BBC is therefore an interesting example of a breed selected for reduced height.

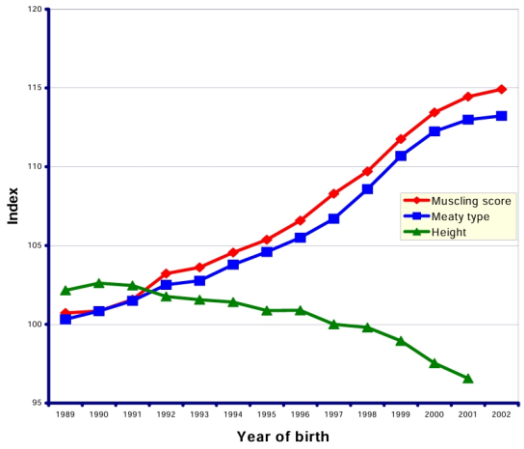


Figure 1.3. Genetic trends for muscular development and height. The figure represents the evolution of breeding values for muscling score, meaty type and height from 1989 to 2002 (copied from Hanset, (2004))

1.1.3 Use of molecular and genomic information in the breeding program

High levels of inbreeding resulting from intensive use of elite sires can lead to outbreaks of recessive genetic defects. This was the case in BBC where Charlier et al. (2008) reported first three genetic defects including congenital muscular dystonia 1 & 2 (CMD1 & CMD2) and crooked tail syndrome (CTS). Fortunately, with the development of genetic arrays and sequencing technologies, it is possible to identify the causative variants and develop genetic tests to manage such recessive alleles in the population. With the help of the Unit of Animal Genomics, such a strategy has been implemented in BBC, where it was decided to exclude all carrier bulls from the breeding program.

The causative variant causing CTS was characterized by Fasquelle et al (2009). Causative variants for other genetic defects have subsequently been identified in BBC, including a splice site variant of the Ring Finger Protein 11 (*RNF11*) gene associated with proportional dwarfism (Sartelet et al., 2012), a missense mutation in the CIC-7 Chloride Channel (*CLCN7*) gene causing osteopetrosis with gingival hamartomas (Sartelet et al., 2014), and a variant in the phosphatidylinositol glycan anchor biosynthesis class H *(PIGH)* gene causing arthrogryposis (Sartelet et al., 2015). Intriguingly, some of these defects segregated at high frequencies in the breed, with 25 and 26 percent of carriers for CTS and dwarfism, respectively (Fasquelle et al., 2009; Sartelet et al., 2012). Such high frequencies may result from the heavy use of popular sires carrying these defects (e.g. Précieux and Galopeur for CTS and dwarfism, respectively) and subsequent high levels of inbreeding. Nevertheless, statistical analyses taking into account the pedigree structure have shown by allele dropping in the true genealogy that the the heavy use of certain sires is not sufficient to explain the observed frequencies and that carriers are positively selected. In addition, Fasquelle et al. (2009) found that carriers of the CTS variants had higher muscular development, consistent with the selective advantage of carriers, whereas the origin of the

selective advantage for the variant causing dwarfism remained unidentified (Sartelet et al., 2012). Using a large sample of 593 progeny tested sires born between 1970 and 2010, Druet et al. (2014a) showed that the variant in mannose receptor C type 2 (*MRC2*) causing the CTS was associated with one of the major QTL for muscular development in BBC, and that the frequency of carriers increased from less than 20% before 1980 to more than 40% in the years 2000. More recently, Gualdrón Duarte et al. (2023, 2020) performed GWAS and genomic predictions in a larger cohort of genotyped cows. Bayesian variable selection models indicated that variants associated with four recessive defects, including variants in *MRC2*, *RNF11*, *ATP2A1* (associated with CMD) and *WWP1*, had large effects on muscular development traits or height. The variant in *WWP1* was previously identified in a reverse genetic screen by Charlier et al. (2016). This variant segregated had relatively high frequency but presented a depletion in homozygotes, suggesting a recessive effect. As for other variants, evidence of balancing selection was also found as the variant was associated with some muscular development traits. Importantly, the study also identified several recessive embryonic lethal variants.

Given the large number of recessive deleterious variants, including the embryonic lethal variant, it was important to genotype all the sires in the population. Therefore, many young sires were genotyped each year and genomic technologies were rapidly adpoted in the breed. This allowed to rapidly reduce frequency of carriers in the population (Table 1.1). However, the exclusion of sires carrying genetic defects reduced the number of bulls in use, with potentially negative effects on genetic diversity and inbreeding levels.

Table 1.1. Evolution of the number of bulls carrying identified genetic defects. The number in brackets indicates the number of bulls genotyped for the corresponding defect (CMD = congenital muscular dystonia; CTS = crooked tail syndrome; DWA = dwarfism; HAM = hamartoma; PG = prolonged gestation; AP = arthrogryposis). The red line in the table indicates the introduction of genetic testing for each defect. Data and table taken from Sartelet (2013).

Year of birth	CMD1 (559)	CMD2 (556)	CTS (539)	DWA (528)	HAM (532)	PG (522)	AP (521)
<2000	17	7	32	17	4	18	1
2000	10	3	41	36	7	11	0
2001	17	4	39	23	8	5	0
2002	14	5	46	28	15	8	5
2003	12	1	32	26	15	12	6
2004	2	5	29	24	11	10	2
2005	2	2	20	26	7	18	5
2006	0	0	13	33	5	5	5
2007	0	0	0	30	14	6	3
2008	0	0	0	6	9	6	3
2009	0	0	0	0	3	3	0
2010	0	0	0	0	0	0	0

The first genomic evaluation in BBC was established in 2016 (Inovéo, 2020). The situation is different from dairy cattle where bulls don't have records for milk production and a long progeny testing was previously required. In beef cattle, many phenotypes of interest have relatively high heritability and own records are available earlier in life, allowing 'visual' selection of young bulls. In addition, in BBC there are fewer progeny tested bulls, not enough to build a large reference population to ensure high reliability. It was therefore decided to build a reference population based on phenotyped cows in a series of reference farms, called genomic farms. Genomic selection was first applied to the linear scoring traits for which convincing results were obtained. The initial goal was to have a reference population of 10,000 genotyped cows with linear scoring phenotypes, an objective that has been clearly achieved since more than such 18,000 cows were available in 2024. Initially the focus was on genotyping animals with as many phenotypes as possible (e.g. adult cows). In parallel, a reference population was also established for carcass traits, including both males and females. After demonstrating the limitations of visual assessment of birth weight, it was decided to properly record birth weight using scales on the genomic farms, and more than 30,000 true birth weights are now available, including 20,000 genotyped animals.

The accumulated datasets allow the genetic architecture of traits recorded in the BBC, such as height or muscular development, to be studied. For example, Gualdrón Duarte et al. (2023) performed multiple trait GWAS at the sequence level and identified several coding variants with large effects. The data also allow to test some genomic prediction selection models or the benefit of using functional annotation, as will be done in the present study.

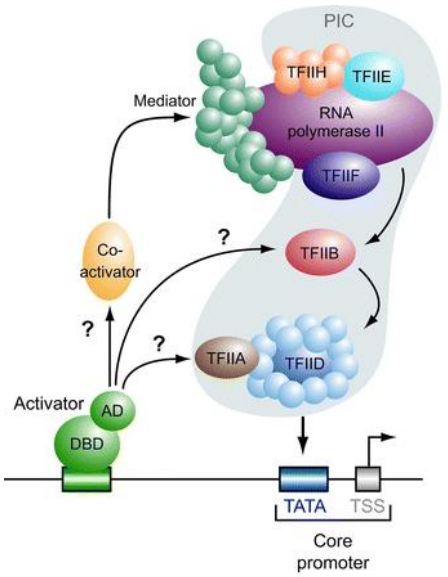
1.2 Functional annotation of the bovine genome

The completion of genome assemblies for many species, including humans, has revolutionized our understanding of genetics and its role in complex traits. Most genes within these newly assembled genomes have been automatically predicted using a combination of ab initio methods and expressed sequence tag (EST) based evidence (Ashurst and Collins, 2003). In addition, the advent of tools such as QTL mapping and GWAS has enabled us to identify genomic regions or variants associated with phenotypes, greatly improving our understanding of the genetic basis of complex traits. However, results from an increasing number of GWAS have shown that many leading SNPs are located outside coding regions (Morova et al., 2022), once thought to be 'junk DNA', expanding our appreciation of the complexity of the genome and its impact on biological processes and phenotypes. Therefore, a comprehensive annotation of the genome sequence is required to gain further insight into how genomic variants, including non-coding ones, are translated into phenotypic differences. This is why several major collaborative efforts such as ENCODE (de Souza, 2012), Roadmap Epigenomics project (Kundaje et al., 2015), the International Human Epigenome Consortium (IHEC) (Stunnenberg et al., 2016) and others have been launched (Forrest et al., 2014; GTEx Consortium, 2020). Unlike the coding regions of the genome, functional elements in non-coding regions often lack easily identifiable features

that would allow their function to be predicted from sequence analysis alone (Birney et al., 2007; Zhen and Andolfatto, 2012). To address this challenge, these consortia have relied on a variety of experimental techniques, including genomic, transcriptomic and epigenomic approaches. In particular, the integration of multiple omics data allows the identification of these functional elements within non-coding regions based on their associations with specific genomic features and their regulatory roles. In parallel with these efforts to understand the functional components of the human genome, several consortia have been established to explore the genomes of livestock and poultry species, including initiatives such as Model Organism ENCyclopedia Of DNA Elements (modENCODE) (Roy et al., 2010), FAANG (Andersson et al., 2015) or GENE-Switch (Vos et al., 2023). Specifically, these consortia refine our understanding of genome functionality by pinpointing gene locations, discovering novel genes and elucidating alternative coding mechanisms (e.g., alternative transcription start sites), as well as uncovering non-coding transcripts such as microRNAs (miRNAs) and long non-coding RNAs (lncRNAs). Most often, these consortia focus on identifying regulatory regions that orchestrate gene expression, such as promoters and enhancers, which are emerging as critical elements alongside protein-coding regions (Andersson et al., 2015). They harbor regulatory elements that interact with transcription factors, chromatin modifying enzymes and other essential components, as highlighted by ENCODE (Pennisi, 2012). In the first part of my thesis, I contributed to the identification of regulatory elements in the bovine genome. Therefore, I start with an overview of these regulatory regions, the methods used to identify them and their importance in complex traits.

1.2.1 What are regulatory elements?

Coding regions account for less than 2% of the human genome, and the remainder was initially considered to be non-functional "junk DNA" (Dunham et al., 2012; Pennisi, 2012), although it has since been shown that non-coding mechanisms play an important role in complex traits. Indeed, the correct spatial and temporal expression of genes involved in a wide range of biological processes, such as differentiation, development and response to stress and stimulation, is central to life (Maston et al., 2006). Perturbations in the regulation of gene expression have also been shown to cause disease and abnormal development (Epstein, 2009; Grant et al., 1996; Van Laere et al., 2003), highlighting their importance. Thanks to the development of sequencing technologies, the ENCODE consortium has identified non-coding regions located in the remaining 98% of the genome that have biochemical activities and serve as landing sites for proteins that coordinate gene expression, both near and far from genes, and defined them as regulatory elements. The published catalogue of identified regulatory elements has allowed significant advances in the understanding of complex traits in humans and mice.



Maston GA, et al. 2006.
Annu. Rev. Genomics Hum. Genet. 7:29–59

Figure 1.4. The assembly of transcriptional machinery. Activators, which are site-specific DNA-binding transcription factors, bind via their DNA-binding domains (DBDs) to specific sequences upstream of the promoter, known as transcription factor binding sites (TFBSs). These activators facilitate the binding of general transcription factors to the promoter region. General transcription factors, including RNA polymerase II and several auxiliary components such as TFIIA, TFIIB and TFIID, interact to form a transcription preinitiation complex (PIC). This complex assembles at the core promoter, guides RNA polymerase II to the transcription start site (TSS) and initiates transcription. Figure adapted from Maston et al. (2006).

Regulatory elements can be defined as non-coding genomic regions that coordinate the level and spatio-temporal dynamics of gene expression. They have specific DNA sequences that are recognised by transcription factors, providing high binding affinities (Boeva, 2016; Nagy and Nagy, 2020). In most cases, the transcription factor binds a specific motif and interacts with so-called co-regulators to form a multi-protein complex that plays a role in the regulatory process (see Figure 1.4). The variation of gene expression levels in different cell types and at different developmental stages reflects the dynamic activity of regulatory elements mediated by their cell-specific transcription factor binding sites (Barral and Déjardin, 2023). Regulatory elements vary in size, most commonly ranging from tens to hundreds of base pairs in length, but rarely exceeding thousands of base pairs. Although regulatory elements have been reported to be located upstream of coding sequences, in introns or even far away from genes, they are significantly enriched in transcription start site (TSS) regions. These regions are responsible for the expression profile of the neighbouring gene(s) and correspond to the so-called proximal promoter. Conversely, studies have reported a depletion of regulatory elements in heterochromatin. The activity of regulatory elements has been shown to be associated with histone modifications (spicuglia and Vanhille, 2012), while chromatin architecture and nucleosomal positioning have been found to influence their accessibility to the transcriptional machinery.

1.2.2 Categories of regulatory elements

Regulatory elements can be classified according to their mechanism (e.g. enhancers versus silencers) and their location (e.g. promoters). They can also be divided into cis- and trans-regulatory elements, affecting nearby and distant genes respectively. Here I will focus on cis-regulatory elements, which can be further categorised into several primary types based on their function, including promoters, enhancers, silencers and insulators (Chatterjee and Ahituv 2017, see Figure 1.5).

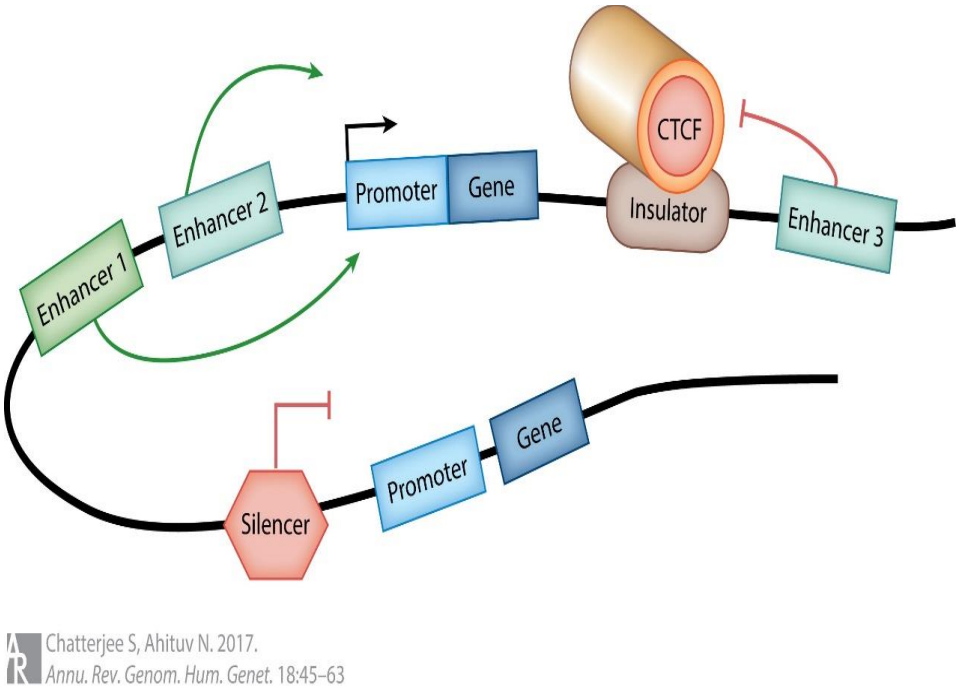


Figure 1.5. The different types of regulatory elements. Promoters directly regulate gene transcription. Enhancers interact with promoters to increase gene expression, while silencers do the same but decrease gene expression. Insulators and CTCF can cooperate to inhibit the function of enhancers (Figure adapted from Chatterjee and Ahituv, 2017).

Promoters are located upstream of genes and are DNA sequences to which RNA polymerase and transcription factors bind to initiate transcription of the corresponding downstream gene. Promoter sequences define the direction of transcription and indicate which strand of DNA is being transcribed. Although the definition of promoter boundaries is not clear, promoters are recognised as regions of approximately 81-1000 bp in length (Xiao et al., 2019) that recruit the transcriptional machinery and serve as its 'landing site'. Their affinity profiles play a crucial role in controlling gene expression, as variation in affinity has been shown to correlate significantly with tissue-specific expression patterns (Molineris et al., 2011). Although the level of conservation of DNA sequences is lower in regulatory regions than in coding regions, mutations in promoters can have a high impact, including on phenotypic morphology (Karim et al., 2011; Zhu et al., 2023).

Enhancers represent a distinct type of regulatory element and are found in non-coding sequences that contain dense clusters of transcription factor binding motifs. They are cis-acting regulatory elements that can be distant from their target genes, up to 1 Mbp (Lettice et al., 2003; Panigrahi and O'Malley, 2021). They increase the expression of a target gene in specific tissues or at specific developmental stages (Blobel et al., 2021; Neyret-Kahn et al., 2023). Enhancers are first recognized and bound by a transcription factor, to trigger the formation of a chromosome loop that allows the enhancer to interact with the promoter, thereby increasing the expression of the target gene. Enhancers have been identified predominantly in intergenic and intronic regions, and less frequently in exons (Panigrahi and O'Malley, 2021). Epigenetic studies have shown that enhancers are almost invariably associated with open chromatin regions and active enhancers are associated with specific histone modifications (Blobel et al., 2021; Sethi et al., 2020). Enhancers have been found to occur in clusters and act in an additive, synergistic or redundant manner to regulate target genes. These clusters are commonly referred to as super-enhancers (Blobel et al., 2021). These super-enhancers are for example significantly associated with tumor-specific genes (Neyret-Kahn et al., 2023). Although reversed genomic orientation of enhancers does not affect their function in gene transcription, subtle modifications in the affinity of enhancers can disrupt the expression of target genes and result in severe and penetrant phenotypes (Lim et al., 2024).

Unlike promoters and enhancers, which positively regulate gene expression, silencers reduce expression by binding to specific transcription factors called repressors. In general, silencers have similar characteristics to those described for enhancers, but with the opposite function (Srinivasan and Atchison, 2004). Silencers are usually located upstream of their gene, but can also be found in introns, exons or in the 3' untranslated region (UTR). Silencers can recruit repressors to compete with activators for the same site, thereby reducing gene transcription or blocking activator binding (Harris et al., 2005; Li and Davie, 2010). Alternatively, these repressors recruited by silencers can reduce the accessibility of a promoter and prevent transcription. This is achieved by altering the chromatin structure, either through histone modifications or chromatin stabilising activities (Harris et al., 2005). Finally, distal silencers can interact with promoters through chromosome loops (Maston et al., 2006).

Finally, insulators are long-range regulatory elements that isolate the independent regulatory domain to prevent environmental noise signals from misleadingly influencing the transcription process (West et al., 2002). An insulator varies in size from 300 bp to 2,000 bp (Allison and Allison, 2008) and regulates the transcription of genes located on the same or a different chromosome that are brought close to the insulator by its looping activity (Allison and Allison, 2008). Two types of insulators have been described, acting either as enhancer blockers or as barriers (West et al., 2002). Insulators that act as enhancer blockers must be located between the enhancer and the promoter (West et al., 2002). Conversely, insulators that act as barriers prevent the expansion of nearby condensed chromatin associated with a silencer. This allows the DNA sequence to have a distinct transcription pattern compared to the nearby region covered by the silencer (West et al., 2002). The CTCF protein is the most

important insulator characterized in vertebrates (Kim et al., 2007; West et al., 2002; Yang and Corces, 2011). CTCF, also known as CCCTC-binding factor, contains an 11-zinc finger (ZF) DNA-binding domain and was identified as an insulator in vertebrates in 1991 (Bell et al., 1999). CTCF also plays an essential role in development and differentiation, is ubiquitously expressed in different cell types and is conserved across species (Kim et al., 2007; Moon et al., 2005). Disregulation of CTCF has been reported to affect histone modification (Fedoriw et al., 2004) and to be associated with a variety of cancers (Filippova et al., 2002, 1998).

1.2.3 Methods to identify regulatory elements

Advances in technologies for the experimental and computational identification of regulatory elements allow for a deeper investigation of their role. In particular, the advent of high-throughput sequencing technologies provides information at the genome-wide level. Reporter gene assays, which use a vector carrying both the sequence of interest and the target gene, are widely used to confirm the regulatory activity of a specific DNA sequence (Kornberg, 1974). However, the focus of my thesis is to generate a catalogue of regulatory elements for the bovine genome, so high-throughput methods are more relevant.

Genomic DNA is hierarchically packed into chromatin, which allows it to be stored in the nucleus but renders most DNA inaccessible (Kornberg, 1974; Richmond and Davey, 2003). However, accessibility and affinity for transcription factors are fundamental features of active regulatory elements. Therefore, chromatin accessibility has been widely used as a marker to identify regulatory elements in the non-coding part of genomes.

Taking advantage of high-throughput sequencing technologies, several approaches have been established to assess chromatin accessibility and identify putative regulatory elements. These include formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE-seq), DNase I hypersensitive sites sequencing (DNase-seq), ATAC-Seq and micrococcal nuclease sequencing (MNase-seq) (see Figure 1.6 for an overview of these methods). One of the major differences between these methods is the distinction between nucleosomal and nucleosome-free regions. In the FAIRE-seq approach, samples are cross-linked with formaldehyde, the DNA is then sheared and a phenol-chloroform extraction is used to recover unfixed DNA fragments in the aqueous phase, corresponding to nucleosome-free DNA, prior to sequencing. As shown in large scale studies, these DNA fragments are significantly enriched for regulatory elements (Giresi et al., 2007; Nagy et al., 2003). DNase-Seq takes advantage of the fact that DNA fragments wrapped by histone octamers are more resistant to deoxyribonuclease I than DNA fragments in open chromatin regions, allowing DNase I hypersensitive sites to be used to identify open chromatin regions. A major advance in the use of DNase-Seq as a method for genome-scale regulatory element detection has been the adaptation of the sequencing technology after DNase I digestion. In recent years, the DNase-Seq approach has been used in large-scale studies, including the ENCODE project (Boyle et al., 2008; Meuleman et al., 2020). ATAC-Seq

also uses a cleavage enzyme, a hyperactive Tn5 transposase, to cut DNA sequences of higher accessibility. The hyperactive Tn5 transposase simultaneously marks the binding site by inserting a sequence adaptor (Figure 1.7). The Tn5 target region is amplified using the inserted primers and the amplified fragments, corresponding to the open chromatin regions, are sequenced. Further details of the procedure are shown in Figure 1.7. Compared to DNase-Seq and FAIRE-Seq, the ATAC-Seq protocol requires fewer and simpler experimental steps and less starting material, making it the most popular approach for open chromatin detection. Thanks to these advantages, ATAC-Seq allows the study of larger samples, covering more tissues or individuals, but also the study of open chromatin at the single cell level. In the MNase-seq approach, the micrococcal nuclease is first used to digest genomic DNA, digesting the naked DNA while leaving undigested the DNA wrapped by histones or other chromatin-bound proteins. The undigested DNA is then sequenced. Thus, unlike the other three approaches, MNase-seq detects nucleosome positions where the interaction of cis-regulatory element and trans-regulatory factor is hindered (Chereji et al., 2019).

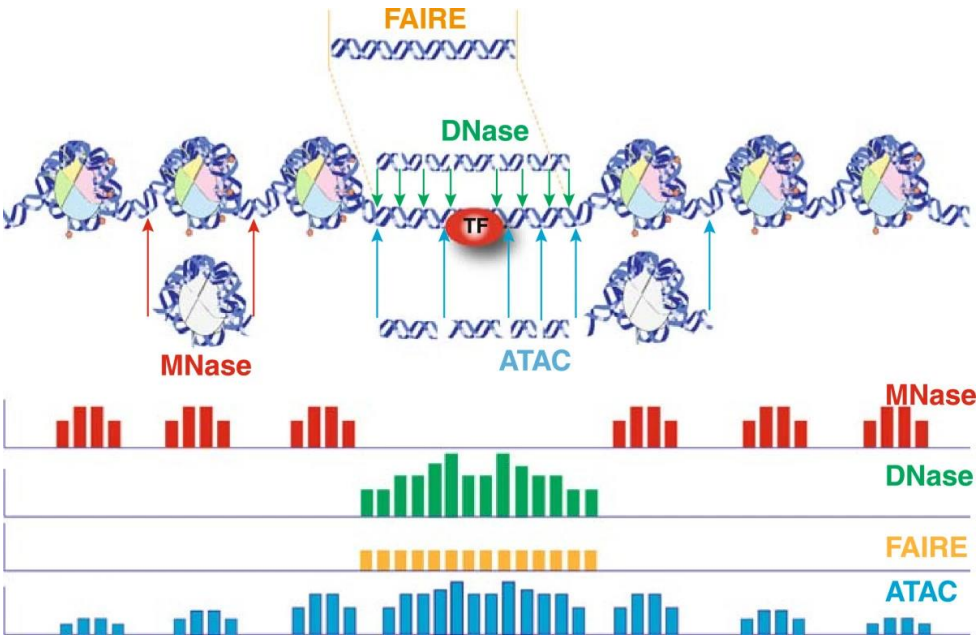


Figure 1.6. The shematic diagram of four experimental based approaches for the detection of chromatin accessibility. ATAC-Seq, DNase-seq, and FAIRE-seq detect open chromatin regions, while MNase-seq identifies condensed chromatin. The arrows indicate the endpoints of the chromatin regions identified by each method. The bars below represent the data signal for each method, reflecting chromatin accessibility (Figure adapted from Tsompana and Buck, 2014).

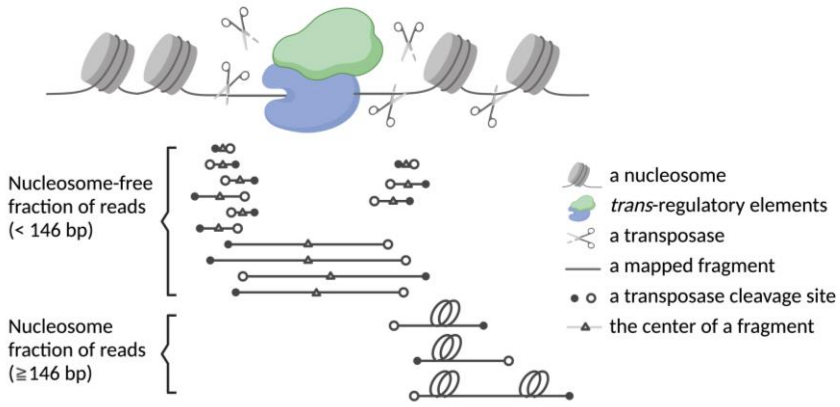


Figure 1.7. The schematic diagram of ATAC-Seq. The transposase cleaves open chromatin (with nucleosomes shown in grey) and inserts sequencing adapters into the DNA. This enables the amplification of regions between two insertion sites, representing open chromatin. The amplified fragments are then sequenced to determine their positions, efficiently delineating the open chromatin regions. Fragments smaller than 146 bp are nucleosome-free, as a single nucleosome is longer than 146 bp.

The activity of the regulatory region is controlled by interactions with transcription factors, nucleosome positioning and DNA and histone modifications. Therefore, the identification of DNA binding sites for transcription factors, the core transcriptional machinery, and different histone modifications provides complementary information to study the regulation of gene expression, especially during development and differentiation (Park, 2009). ChIP sequencing, also known as ChIP-seq, was developed by combining chromatin immunoprecipitation assays with massively parallel sequencing (Robertson et al., 2007). ChIP-seq is the most universal approach to identify DNA sequences bound by a specific protein, such as transcription factors or other chromatin-associated proteins with different epigenetic modifications (Park, 2009). In ChIP-seq, antibodies specific to the protein of interest are first used to enrich the DNA fragment bound by that protein. The precipitated DNA is then subjected to high-throughput sequencing, allowing genome-wide mapping of the region bound by that protein. Since its development, an increasing number of studies have used ChIP-seq to decipher the transcriptional binding map. Among these, the identification of CTCF binding sites in different tissues and species has been of particular interest due to its critical role in regulation (Holwerda and de Laat, 2013; Oomen et al., 2019; Phillips and Corces, 2009).

In addition, epigenetic modifications including DNA methylation and histone post-translational modifications such as acetylation, SUMOylation, phosphorylation, methylation and ubiquitination have been reported to regulate gene expression. These modifications don't alter the DNA sequence, but affect histone-DNA interactions, resulting in a change in chromatin state (Espinoza Pereira et al., 2023; Kimura, 2013). In mammals, DNA methylation predominantly targets cytosine nucleotides followed by guanine nucleotides, known as CpG sites, with up to 75% of CpGs in the genome being methylated (Tost, 2009). However, CpG sites within CpG islands, which are regions enriched for regulatory elements, tend to have lower levels of methylation (Moore et al., 2013). Among the many techniques

available to detect DNA methylation, bisulfite sequencing is the most widely used and efficient method (Li and Tollefsbol, 2011). It allows precise profiling and quantification of DNA methylation at single base pair resolution across the genome and has been used in the Epigenomics Roadmap project. The interplay between histones and the genome is essential for nucleosome positioning and chromatin organisation (Parmar and Padinhateeri, 2020), and thus critical for transcriptional regulation by influencing DNA accessibility. Modifications to histones can alter these interactions, thereby dynamically shaping the precise programmes that govern nucleosome positioning. The role of histone modification varies depending on the type of modification and the specific amino acid residues, resulting in different mechanisms. For example, histone H3 lysine 4 methylation (H3K4me3) and histone H3 lysine 27 acetylation (H3K27ac) promote gene expression, whereas histone H3 lysine 9 trimethylation (H3K9me3) represses gene expression (Kimura, 2013).

Using ChIP-seq with antibodies targeting specific histone modifications, it is possible to accurately detect genomic regions associated with these signatures and to reveal the locations of histones bearing these modifications along the DNA sequence. A comprehensive histone modification map of 127 cell lines and tissues covering the major cell lineages in the human body has been generated (Kundaje et al., 2015). In addition, by integrating histone modifications with genome accessibility data, we can delineate the chromatin state of regulatory elements and identify active elements (Boix et al., 2021; Hoffman et al., 2013; Kern et al., 2021). In the Roadmap Epigenomics project, a hidden Markov model, called the ChromHMM model, has been developed to identify 15 core chromatin states that capture primary interactions between epigenetic marks, including Active TSS, Flanking Active TSS, Strong Transcription, Genic Enhancers and more (Kundaje et al., 2015).

Although the experimental techniques available for identifying regulatory elements have been greatly improved by high-throughput sequencing technologies, they remain labour-intensive and expensive, and are therefore mainly accessible to model organisms. Even in these species, it remains difficult to cover all developmental stages and cell types and to work on many individuals. Bioinformatic approaches therefore provide an alternative and complementary approach to identifying regulatory elements. Bioinformatic approaches could also improve experimental results by correcting some of the environmental "noise". However, as these bioinformatic approaches are not the focus of my thesis, I will only briefly summarise the most common methods:

1. Identification of TF binding motifs: Short DNA sequences that are specifically recognised by TFs are referred to as binding motifs (Leporcq et al., 2020). The most commonly used methods for finding TF binding sites are based on position weight matrices (PWMs) defined for a given set of known motifs (Staden, 1984). A weight matrix records the score of four nucleotides at each position of a given motif, allowing a quantitative assessment of the potential of sequences of interest to be a motif (Staden, 1984). Nowadays, the increasing availability of PWMs for known motifs in public

databases such as JASPAR (Sandelin et al., 2004) and the TRANSFAC database (Wingender et al., 2000) greatly facilitates the prediction of TF binding sites. In addition, PWMs of de novo motifs have also been derived from de novo motif discovery programmes, for example from ChIP-seq data (Leporcq et al., 2020).

2. Identification of regulatory elements using phylogenetic footprinting: Phylogenetic footprinting, another experimental-free technology, identifies regulatory elements by analysing the conservation of orthologous regulatory regions across species (Tagle et al., 1988). Regulatory regions are thought to be more conserved than other non-coding genomes because of their functional importance in controlling gene expression (Wasserman and Sandelin, 2004). The phylogenetic footprint hypothesis relies on the assumption that regulatory mechanisms for a given gene remain consistent across species. Various methods embodying this approach have been developed (Bailey and Elkan, 1995; Blanchette and Tompa, 2002; Hertz and Stormo, 1999) and have been extensively used to reveal the regulatory elements of genes (Leung et al., 2000; Loots et al., 2000; Manen et al., 1994). As the number and quality of reference genome assemblies increase, they serve as a valuable resource for inferring regulatory elements on a genome-wide scale. Christmas et al. (2023) have predicted cross-species cis-regulatory elements at the tissue and cell type level by aligning hundreds of placental mammalian genomes.

1.2.4 Identification of regulatory variants

To fully understand the regulatory architecture and gain insight into how regulatory elements contribute to complex diseases and traits, projects have been initiated to generate additional information that complements the identification of regulatory elements. One prominent initiative is the GTEx project (GTEx Consortium, 2020). GTEx aims to elucidate how genetic variants affect gene expression in different tissues and individuals, providing critical insights into the role of regulatory variants. GTEx has investigated the effect of genetic variants on gene expression using eQTL analysis. This analysis helps to identify variants in the regulatory element that are associated with differences in gene expression levels between individuals. By studying eQTLs in different tissues and populations, GTEx is improving our understanding of the regulatory landscape that controls gene expression. In addition, GTEx has been extended to different species (Guan et al., 2023; Liu et al., 2022; Teng et al., 2024), allowing comparative analyses to reveal evolutionary conservation and divergence in regulatory mechanisms, facilitating the identification of conserved regulatory elements and their functional significance across species.

In the GTEx study, two primary types of eQTL signals were identified: cis-eQTL and trans-eQTL (GTEx Consortium, 2015). Cis-eQTL refer to genetic variants located in proximal regulatory regions that influence the expression of nearby genes and are the predominant type of eQTL observed.

These variants are typically located within cis-regulatory elements. Conversely, trans-eQTLs refer to variants that affect the expression levels of multiple genes located distally, and many of these eQTLs are associated with transcription factors. It's worth noting that splicing eQTLs are excluded from our analysis as they fall outside the scope of this study. Importantly, the majority of eQTLs operate across multiple tissues and regulatory variants may disrupt the expression of more than one gene. Cis-eQTLs generally show larger effect sizes on gene expression compared to trans-eQTLs and have a higher replication rate across different tissue types (Võsa et al., 2021).

1.2.5 The advances of functional annotation in cattle

In cattle, efforts have also been made to improve the functional annotation of the reference genome since its completion (Pareek et al., 2011). These efforts have focused on gaining deeper insights into the regulation and biological functions of various genes, particularly their roles in the intricate biological processes underlying differentiation and development. As mentioned above, the advent of improved sequencing technologies has led to the generation of vast amounts of data over the past few decades. Among these, two prominent consortia, namely FAANG, launched in 2015 (Andersson et al., 2015), and BovReg, part of Euro FAANG, launched in 2020 (Moreira et al., 2022), have made significant contributions to advancing our understanding of the bovine genome and its functional elements. Using data from the FAANG project, researchers have extensively studied chromatin accessibility in different tissues and under different conditions, including different breeds, and compared them with those of other domesticated animals, revealing conserved regulatory patterns (Fang et al., 2019; Foissac et al., 2019; Kern et al., 2021). The BovReg project, similar to the ENCODE project in humans, aims to create a comprehensive regulatory map of the bovine genome across tissues, ages and populations by integrating data on gene expression, epigenetic modifications and chromatin accessibility (Moreira et al., 2022). Recently, a new catalogue of transcription start sites across different tissues has been annotated by BovReg (Salavati et al., 2023). In addition, other studies have investigated more specific aspects such as chromatin accessibility in Indicine cattle (Alexandre et al., 2021), the immune epigenome of different breeds (Powell et al., 2023) and chromatin dynamics in different tissues (Gao et al., 2022a; Halstead et al., 2020b). In line with the human GTEx project, the cattle GTEx (cGTEx) initiative has published the results of its pilot study (Liu et al., 2022). This includes a catalog of regulatory variants that provides insights into how genetic variation affects gene expression in different bovine tissues (Liu et al., 2022). These studies are making a significant contribution to our knowledge of specific regulatory elements and their dynamics in bovine biology. Despite extensive efforts, the annotation of the bovine genome remains less advanced than that of the human genome. This lack of thorough annotation poses a challenge in identifying and understanding the functional elements that contribute to complex traits.

1.3 Heritability partitioning using functional annotation

The proportion of phenotypic variance attributable to genetic variation is termed heritability and is commonly estimated using pedigree or genomic information (Visscher et al., 2008). The heritability captured by genotyped variants, mainly variants in LD with causal variants and, less frequently, some causal variants, is referred to as SNP heritability (Yang et al., 2017). Heritability estimation is essential for genetic evaluation and genomic prediction. Despite the widespread use of GWAS to identify variants associated with complex traits, the genetic variance explained by the identified SNPs remains far below the total heritability for most traits (Eichler et al., 2010; Yang et al., 2011b). To better estimate SNP heritability and account for the collective contribution of multiple genetic variants, several methods have been developed to fit all genetic variants simultaneously, such as the genomic restricted maximum likelihood (GREML) approach (Yang et al., 2011a) and BayesR (Moser et al., 2015). Extensions of these methods allow the contribution of different categories of variants to the SNP heritability to be estimated, potentially improving our understanding of the genetic architecture of complex traits. For instance, SNP categories could be defined based on their location (e.g. chromosomes), minor allele frequencies (MAF), levels of LD with other variants, GC content or functional annotation (Gusev et al., 2014; Lee et al., 2012; Yang et al., 2011b). The continuing increase in the number of whole-genome sequenced individuals, coupled with improvements in the functional annotation of livestock genomes, provides an unprecedented opportunity to apply such approaches to partition the heritability of important agronomic traits.

1.3.1 Method used for heritability partitioning

To date, three main approaches have been developed to perform heritability partitioning: methods based on GREML (Yang et al., 2011a), Bayesian mixture models (Patxot et al., 2021), and so-called LD score regressions (Finucane et al., 2015) (the LD score associated with a SNP is defined as the sum of the LD levels of that SNP with other SNPs in the region). Of these, GREML and Bayesian mixture models allow the use of both individual-level data and summary statistics, whereas LD score regression is primarily restricted to the use of summary statistics and is commonly used in human genetic research.

GREML is an approach proposed by Yang et al. (2011a) that uses REML to estimate variance components from a linear mixed model (LMM) containing one or more polygenic terms associated with a genomic relationship matrix (GRM). Following common practice (e.g., Mrode), I use GBLUP to denote such an LMM with a single polygenic term modelled using genomic information, although Best Linear Unbiased Prediction (BLUP) refers to the estimation of breeding values with LMM in general (e.g., with an animal model). Thus, in my thesis, GBLUP will refer to a model, while GREML will refer to the estimation by REML of the variance components of a GBLUP model (or other genomic LMMs). The standard GBLUP models assume a priori that all SNPs contribute equally to the genetic variance

(Meuwissen et al., 2001) or according to their respective allele frequencies (see below). This model can be written as:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{g} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is the vector of individual phenotypes, $\mathbf{1}\boldsymbol{\mu}$ is the intercept term (i.e. the mean effect), $\boldsymbol{g}$ is the vector of individual polygenic effects, and $\boldsymbol{e}$ is the vector of individual random error terms. The polygenic effects are normally distributed, $\boldsymbol{g} \sim N(0, \mathbf{G}\sigma_g^2)$ where **G** is the genomic relationship matrix (GRM). The calculation of **G** relies on the *n* x *m* genetic dosage matrix **M**, which indicates how many alternate alleles the *n* individuals carry at each of the *m* markers. The dosage of individual *i* and marker *j*, m_{ij} is typically 0, 1 or 2 when relying on genotypes, but other values are possible when working with genotype probabilities. Note also that different allele coding can be used, such as reference versus alternate allele, ancestral versus derived allele, major versus minor allele or A versus B allele. To construct the GRM, the centred dosage matrix **Z** or the centred and scaled dosage matrix **X** are generally used. The allele frequencies are used to centre the genotypes as $z_{ij} = m_{ij} - 2f_j$, where f_j is the allele frequency of the alternate allele at marker *j*. The scaling of individual centred dosages is achieved as:

$$x_{ij} = \frac{m_{ij} - 2f_j}{\sqrt{2f_j(1-f_j)}}.$$

Using the centred dosages, the GRM is estimated as:

$$\mathbf{G} = \frac{\mathbf{ZZ'}}{\sum_{j=1}^{n} 2f_j(1-f_j)},$$

where the denominator is used to scale **G** like the additive relationship matrix **A** (VanRaden, 2008). With the scaled and centred dosages, the GRM is obtained as:

$$\mathbf{G} = \frac{\mathbf{XX'}}{n}.$$

Although these rules have been described in several studies (Speed and Balding, 2015; VanRaden, 2008; Yang et al., 2011a), the centred GRM is often referred to as "VanRaden1", while the scaled and centred GRM is referred to as "Yang" (although it also corresponds to "VanRaden 2"). In the first case, the effect sizes of markers are independent of their allele frequencies, and common alleles consequently have higher contributions to genetic variance. In the second case, standardisation ensures that each marker has an equal expected contribution to genetic variance, and rare alleles will have large effect sizes.

For heritability partitioning, the LMM is extended to multiple polygenic terms, one per annotation group (e.g., different categories of genetic variants classified according to their allele frequency, function, etc.), allowing the genetic variance associated with each of these groups to be estimated:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \sum_{s=1}^{S} \boldsymbol{g}_s + \boldsymbol{e},$$

where $\boldsymbol{g}_s$ is the vector of individual polygenic effects associated with annotation group *s*, *S* is the total number of fitted annotation groups. Each polygenic component is normally distributed, $\boldsymbol{g}_s \sim N(0, \boldsymbol{G}_s\sigma_s^2)$ and has its own GRM $\mathbf{G}_s$, computed using only the variants present in category *s*, and its own variance σ_s^2. In this model, the additive polygenic variance, σ_g^2, is equal to the sum of the variances associated with each annotation groups:

$$\sigma_g^2 = \sum_{s=1}^{S} \sigma_s^2.$$

The contribution of annotation group *s* to the genetic variance, called %SNP heritability, is estimated as:

$$\%h_s^2 = \frac{\sigma_s^2}{\sigma_g^2}.$$

As described above, different rules have been proposed to construct the GRM, the most commonly used being the so-called "VanRaden1" and "Yang" rules. These rules correspond to different assumptions about the relationship between allele frequency and SNP effect size. However, the true properties of causal variants in terms of MAF or LD levels remain unknown and may differ from the assumptions made in GREML. This has been reported to affect heritability estimation (Speed et al., 2012) and therefore several GREML-based approaches have been developed to allow for more general assumptions and greater flexibility, such as the GREML-LDMS (Yang et al., 2015) and LDAK (Speed et al., 2012). In the GREML-LDMS approach (Yang et al., 2015), SNPs are assigned to multiple annotation groups based on their LD scores and MAF. Each category has a different distribution of SNP effects, allowing more flexibility in the relationship between LD and MAF and SNP effect sizes. The model thus accounts for LD and MAF stratification (LDMS). In LDAK, LD and MAF effects are adjusted by weighting the variants used to calculate the GRM based on their LD score and allele frequency (Zhang et al., 2021). These derived models show reduced bias and improved robustness in heritability estimation. Consequently, LD and MAF corrections are commonly used in partitioning approaches. However, the increased number of variance components to be estimated poses

computational challenges, particularly for the largest cohorts. For example, the average information-REML (AI-REML) algorithm allows faster convergence (Gilmour et al., 1995), but can have convergence problems when parameters are close to the boundary of the parameter space (e.g. negative variances or correlations > 1). Expectation–maximization-REML (EM-REML) is more stable and can converge even in scenarios where AI-REML fails, but at a slower rate (Misztal, 2008) and with a potential risk of becoming trapped in local maxima. To address the computational challenges posed by large datasets, RHE-mc (based on Haseman-Elston regression) (Pazokitoroudi et al., 2020) and BOLT-REML (Loh et al., 2015a) have recently been developed. REML-based approaches are unable to efficiently handle large datasets, making methods like RHE-mc and BOLT-REML more practical for big dataset in future. These methods are faster and capable of analyzing large datasets, which can lead to higher accuracy due to the increased number of individuals being analyzed.

Bayesian mixture models allow the construction of more complex and realistic models of heritability partitioning by (1) estimating parameters using non-infinitesimal hypotheses that allow only a subset of SNPs to have genetic effects, (2) incorporating flexible SNP effects with non-Gaussian distributions that include both small and large effects, and (3) allowing the specification of priors for all parameters in the model. A variety of Bayesian models are commonly used in genomic prediction, such as BayesB (Meuwissen et al., 2001), BayesCπ (Habier et al., 2011) and BayesR (Erbe et al., 2012). These models have been reported to improve the accuracy of genomic prediction over GBLUP due to their ability to better define SNP effects and to accurately incorporate the number of SNPs with a non-zero effect into the model (Moser et al., 2015). However, only a few of these approaches have been extended to accommodate different annotation groups, as required for heritability partitioning, including BayesRC (MacLeod et al., 2016), BayesRCO (BayesRCπ and BayesRC+) (Mollandin et al., 2022), and BayesRR-RC (Patxot et al., 2021). These methods are derived from BayesR and use different effect sizes (large, medium, small and zero) to better represent the realistic distribution of SNP effects. Of these, only BayesRR-RC has the ability to directly estimate the variance contributed by each SNP group. In BayesRR-RC, phenotypes are modelled as:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \sum_{s=1}^{S} \mathbf{X}_s \boldsymbol{\beta}_s + \boldsymbol{e},$$

where $\mathbf{X}_s$ is the matrix of centred and scaled genotypes for markers in category *s* and $\boldsymbol{\beta}_s$ is the vector of marker effects for category *s*. These are modelled as a mixture of null effects (spike probability at zero) and *L* Gaussian distributions:

$$\beta_{s_j} \sim \pi_{0_s}\delta_0 + \pi_{1_s}N\left(0, \sigma^2_{1_s}\right) + \pi_{2_s}N\left(0, \sigma^2_{2_s}\right) + \ldots + \pi_{L_s}N\left(0, \sigma^2_{L_s}\right),$$

where j is the marker index, δ_0 is a discrete probability mass at 0, L is the number of Gaussian distributions in the mixture, $\{\pi_{0_s}, \pi_{1_s}, \pi_{2_s}, \dots, \pi_{L_s}\}$ are the mixture proportions for annotation group s, $\{\sigma^2_{1_s}, \sigma^2_{2_s}, \dots, \sigma^2_{L_s}\}$ are the mixture variances for group s and correspond to predefined proportions of σ^2_s, the variance explained by the group estimated directly from the data. The proportion of heritability associated with each group is then estimated as:

$$\%h^2_s = \frac{\sigma^2_s}{\sum_{s=1}^{S} \sigma^2_s}.$$

Compared to BayesRR-RC, BayesRC lacks the explicit estimation of the variance for each annotation group in the model. Indeed, the group-specific variances σ^2_s are replaced by the total additive genetic variance σ^2_g. Therefore, Xiang et al. (2023) used an indirect approach to compute σ^2_s using the following formula:

$$\sigma^2_s = 0.0001\ \sigma^2_g \pi_{1_s} + 0.001\ \sigma^2_g \pi_{2_s} + 0.01\ \sigma^2_g \pi_{3_s},$$

where 0.0001, 0.001 and 0.01 correspond to the predefined proportions of σ^2_g associated with the three distributions he fitted per annotation group in his model.

Unlike BayesRC and BayesRR-RC, which explicitly assign each SNP to a single annotation group, a process that can be subjective, BayesRCO (Mollandin et al., 2022) allows variants to have multiple annotation groups. The probability of a SNP to belong to each group is modelled as an additional variable and estimated in parallel with other parameters.

While Bayesian-based methods offer a better defined variance scheme, GREML remains prevalent in many analyses on individual-level data (i.e., with individual phenotypes for each genotyped individual) due to its faster computational times compared to Bayesian mixture approaches, which often require longer sampling chains to reach equilibrium. Two recently developed software GMRM (Patxot et al., 2021) and BayesR3 (Breen et al., 2022) use a different strategy to improve speed, making heritability partitioning at the sequence level feasible. Conversely, convergence problems become more frequent with GREML when too many annotation groups are fitted simultaneously.

In some cases, however, genetic analyses rely on summary statistics rather than individual-level data. LD score regression (Bulik-Sullivan et al., 2015) (or LDSC) is an approach that can be used to estimate heritability and perform heritability partitioning with summary statistics. It is based on the principle that in association studies, the estimated effect of a genetic variant results from the combination of its own effect and that of other variants that are in LD, but also from other factors such as population stratification. LDSC distinguishes between inflation due to polygenic effects and the consequences of

cryptic relatedness or population stratification by modelling the relationships between test statistics and LD scores. More specifically, the expected association statistic for variant j has a χ^2 distribution (Bulik-Sullivan et al., 2015):

$$E\left[\chi_j^2 \middle| l_j\right] = nh^2 l_j/m + na + 1,$$

where n is the total sample size, m is the number of markers, $h^2$is the heritability, $l_j = \sum_k r_{jk}^2$ is the LD Score of variant j(r_{jk}^2 is the LD between markers j and k), a is the contribution of confounding biases such as stratification and cryptic relatedness to the test statistic.

This approach can be extended to partition heritability according to functional annotation (Finucane et al., 2015). In this case, the test statistic is modelled as:

$$E\left[\chi_j^2\right] = n\sum_s \tau_s l(j,s) + na + 1,$$

where s is the annotation group index, $l(j,s)$ is the LD Score of marker j with respect to category s (i.e. $l(j,s) = \sum_{k\in s} r_{jk}^2$), τ_s is the per-SNP contribution to heritability of annotation group s.

Each of the three approaches has its own advantages. A key feature of LDSC is that only a simple regression is required to estimate τ_s, significantly reducing computational time and resources compared to GREML and Bayesian mixture models, even for large cohorts. This allows analysis based on large numbers of individuals and annotation groups. For example, Finucane et al. (2015) successfully implemented LDSC to perform heritability partitioning for 17 complex diseases and traits across 53 annotation categories, which would be computationally intractable with GREML. They also showed that including more annotation groups improves the accuracy of heritability partitioning. In addition, stratified LDSC offers greater flexibility in group specification, allowing a variant to be assigned to multiple groups, whereas GREML and Bayesian approaches, with the exception of BayesRCO, only allow one annotation group for each SNP. It is worth noting that, unlike these GREML and Bayesian mixture models, LDSC can distinguish population structure from polygenic effects, resulting in improved heritability estimation in the presence of population stratification. However, LDSC only works with summary statistics, and if the genotype data of the individuals used to generate these statistics are not available, it is crucial that the individuals used to calculate LD scores have similar LD patterns to those used in the GWAS. Although LDSC has proven effective in human genetics and has been further improved over the years, it is rarely used in animal breeding. Importantly, LDSC has been reported to be inferior to approaches based on individual-level data in both humans (Patxot et al., 2021) and cattle (Xiang et al., 2023). For example, in the study by Patxot et al. (2021), heritability partitioning using LDSC was found to be less accurate than BayesRR-RC and Bolt-REML, which use individual-level data. GREML assigns the same variance parameter to all SNPs in the same annotation group,

whereas Bayesian mixture models offer greater flexibility at the expense of more parameters to estimate. GREML generally results in faster computation, but the number of annotation groups that can be fitted simultaneously remains limited. Indeed, partitioning with a large number of annotation groups dramatically increases the computational time and often results in convergence problems (Finucane et al., 2015). Bayesian mixture models are generally slower due to the time-consuming sampling process required for each SNP. However, Bayesian methods provide additional posterior inclusion probability (PIP) information and SNP effects for each SNP, which can be used for fine mapping and provide higher resolution than standard LMM-based GWAS. Finally, BayesRR-RC showed comparable accuracy in heritability partitioning to Bolt-REML and higher accuracy than RHEmc in the simulation study by Paxtot et al. (2021).

1.3.2 Applications of heritability partitioning in humans and livestock populations

The development of these techniques has allowed the contribution of different genomic features and functional elements to genetic variance to be studied, providing a better understanding of complex traits in humans, model organisms and livestock species. The first partitioning approaches investigated the contribution of different chromosomes to genetic heritability. For example, Yang et al. (2011b) used GREML to partition the heritability of human height and body mass index (BMI) across chromosomes and found that the contribution from each chromosome was correlated with its length, suggesting highly polygenic traits. Similar studies were later applied to other traits (Loh et al., 2015a) and in livestock (Bhuiyan et al., 2018; Jensen et al., 2012; Pimentel et al., 2011). For example, extreme polygenicity has been observed for schizophrenia (Loh et al., 2015a). In contrast, a few traits showed a different architecture, where a single gene was associated with most of the genetic variation, such as the ABO locus for plasma von Willebrand factor (vWF) levels (Yang et al., 2011b). Next, partitioning approaches were used with annotation group associated with LD score or MAF categories (Bhuiyan et al., 2018; Yang et al., 2015). In humans, the use of such LDMS approaches was shown to be necessary to obtain unbiased estimates of heritability (Yang et al., 2015). In addition, results from GREML-LDMS suggest that the heritability of human height is enriched for variants with a MAF < 0.1.

With the increased functional information available, partitioning approaches have been used to estimate the contribution of different functional classes to complex traits. These studies have mostly been carried out in humans. For example, Gusev et al. (2014) used GREML to estimate the contribution of six functional categories, including coding, UTR, promoter, DNaseI hypersensitivity sites (DHS), intronic and intergenic, to 11 common diseases. Their analysis showed that variants in coding regions had the highest levels of per-SNP heritability (i.e. the proportion of genetic variance explained by a SNP in that category), while regulatory elements explained the largest proportion of genetic variation (their collective contribution). Furthermore, in another study using the LDSC approach to unravel the genetic architecture of 49 disease and complex traits (van de Geijn et al., 2020), transcription binding sites within regulatory regions were found to make a substantial contribution to heritability. Using the same

approach and data set, Hujoel et al. (2019) showed that conserved regulatory elements had higher enrichment levels compared to broad regulatory elements. The use of Bayesian mixture models to partition heritability is more recent. For example, Patxot et al. (2021) used BayesRR-RC to investigate the genetic architecture of four complex traits including cardiovascular disease (CAD) outcomes, type 2 diabetes (T2D), BMI and height. Their analysis revealed a slightly different genetic architecture, with a large proportion of heritability associated with intronic, exonic and distal regulatory regions, and less than 10% associated with proximal regulatory regions. Their study also confirmed that common variants explain a substantial proportion of heritability. The variant effect sizes from each category showed a parallel trend to the proportion of heritability associated with the categories (e.g. intronic and exonic regions had larger effect sizes). Finally, partitioning approaches based on summary statistics allow more annotation groups to be included in the model. For example, a baseline model with more than 50 annotation groups has been used in several human studies (Finucane et al., 2015; Speed et al., 2020; Speed and Balding, 2019; Zheng et al., 2024). These analyses revealed substantial heritability enrichment levels in conserved regions for several complex traits (Finucane et al. 2015). Surprisingly, this enrichment exceeded that of coding variants, although Speed et al. (2019) pointed out a possible overcorrection.

Similar studies have been carried out in cattle, although functional annotation is not as complete and accurate as in humans. First, genome partitioning of genetic variation per chromosome with GREML showed a weak correlation between length and associated variance for traits such as milk production and composition (Pimentel et al., 2011) and fitness (Jensen et al., 2012) recorded in dairy cattle. This suggests an uneven distribution of QTL and the presence of large effect variants, which is not consistent with findings in humans. Similar results have been observed for carcass traits in beef cattle (Bhuiyan et al., 2018; Niu et al., 2021). Functional heritability partitioning has also been applied to livestock species. For example, Koufariotis et al. (2014) performed one of the first studies in dairy and beef cattle. However, they mainly used high-density genotyping arrays and in silico prediction of functional classes (as obtained with the Variant Effect Predictor - VEP (McLaren et al., 2016)). They found that variants in coding and regulatory regions contributed significantly more to heritability, consistent with findings in humans (Gusev et al., 2014). To do this, they estimated the contribution of only one category at a time by fitting a GREML with two annotation groups (the target category versus the rest of the genome). In fact, this strategy has been used in most studies carried out in livestock species to estimate the contribution of different genomic features, both in cattle (Edwards et al., 2015; Lingzhao et al., 2017) and in other species such as pigs (Sarup et al., 2016; Ye et al., 2020), and is the core of the so-called Genomic Feature BLUP (GFBLUP) (Edwards et al., 2016). More recently, Xiang et al. (2023) used a Bayesian mixture model to highlight the important role of regulatory variants in complex traits in cattle. Compared to many of the previous studies, they relied on more advanced functional information, including information from the cattle GTEx (cGTEx) database. Overall, the results from different studies in cattle have been highly variable, with the estimated contribution of each class varying

significantly from trait to trait (Xiang et al., 2023). Sometimes unexpected results have been reported, such as synonymous variants explaining a larger proportion of the genetic variance of carcass traits than non-synonymous variants (Bhuiyan et al., 2018). Finally, the LDSC approach has been used very rarely in livestock species, where it has been reported to achieve low accuracy (Xiang et al., 2023).

It's worth noting that these analyses were carried out using methods developed in the human genetic community and tested mainly on human data. Although these methods have been widely used in livestock species, their accuracy has not been fully validated in this context. To my knowledge, the only study in livestock that has evaluated the accuracy of GREML for estimating the variance associated with different annotation groups was performed by Cai et al. (2022). The study focused only on LDMS stratification in a relatively small dataset (~2000 individuals), not at the sequence level, and only with a two variance component approach (not by fitting all annotation groups simultaneously). However, livestock populations have very different characteristics from human data at the genomic level. For example, they typically exhibit characteristics such as long distance LD, small effective population size, and high levels of inbreeding and relatedness between individuals. Such characteristics can significantly affect the efficiency and accuracy of heritability partitioning methods. Therefore, a comprehensive evaluation of the properties of these methods in livestock populations would be valuable.

1.4 Using biological priors in genomic selection models

1.4.1 Introduction to genomic selection

The concept of GS, originally introduced by Meuwissen et al. (2001), is based on the use of high-density genotyping to capture all QTLs across the genome. GS is designed to simultaneously capture the effects of all causal variants distributed across the genome, including those with small effects that don't reach significance levels in GWAS. Compared to conventional breeding approaches based on phenotypic and pedigree data, or to Marker-Assisted Selection (MAS), which uses a limited number of markers to capture only large effects - QTL or genome-wide significant associations - GS offers the potential to achieve significantly higher prediction accuracies for complex traits in plant and animal breeding (García-Ruiz et al., 2016; Georges et al., 2019; Meuwissen et al., 2001; Zhao et al., 2012). In addition, GS can also increase genetic progress by reducing the generation interval. In practice, the advent of low-cost commercial genotyping arrays, such as the SNP 50K bovine chip, has greatly accelerated the application of GS, making large-scale genotyping feasible. Accordingly, the adoption of GS has become widespread among breeding companies, particularly in cattle, as a replacement for laborious and costly progeny testing (de Koning, 2016; Georges et al., 2019). In the US dairy industry, García-Ruiz et al. (2016). reported a twofold increase in annual genetic gains for milk, fat and protein yields and a significant reduction in the generation interval for selection of sires of bulls and cows when comparing periods before and after the introduction of genomic selection in 2010 (García-Ruiz et al., 2016). Similar trends were observed for the Montbeliarde, Normande and Holstein breeds in France

(Doublet et al., 2019). The trend reported for Holstein bulls was accompanied by a marked reduction in genetic diversity, exceeding the acceptable rate of inbreeding defined in the FAO guidelines (Doublet et al., 2019). The decrease in genetic diversity could be compensated by increasing the number of new bulls and by optimising the use of genomic information in breeding programmes, e.g. to balance the level of inbreeding and genetic gain (Doublet et al., 2019). The implementation of GS in the beef industry presents unique challenges compared to dairy cattle, such as the ability to directly observe phenotypes in young males or the lower use of artificial insemination. In addition, the presence of multiple breeds results in smaller population sizes per breed, which complicates prediction efforts. Compared to dairy cattle, the reference population of genotyped animals remains small, especially compared to the Holstein breed, and phenotypes may also be of lower quality in beef breeds (Meuwissen et al., 2016; Van Eenennaam et al., 2014). Although the use of a multi-breed reference population strategy has shown marginal improvements in accuracy (when the same phenotypes are recorded in the different breeds), the presence of shared QTLs becomes crucial when performing such cross-breed GS. In addition, the accuracy of multibreed prediction decreases significantly when using a 50K SNP panel due to differences in LD phases between breeds (Meuwissen et al., 2016). Although GS has been adopted in other species such as pigs, poultry, sheep or fish (Georges et al., 2019; Ibáñez-Escriche et al., 2014; Meuwissen et al., 2016), this adoption is not as widespread as in cattle. Finally, similar techniques, called polygenic risk scores, are being investigated in human genetics, for example to predict disease risk (de los Campos et al., 2010).

1.4.2 Statistical approaches used in genomic selection

Some of the models used for GS have already been introduced because they are similar to approaches used for heritability partitioning. The SNP-BLUP and GBLUP, two equivalent LMM models as previously shown (see for instance Goddard, 2009; Strandén and Garrick, 2009; Strandén and Christensen, 2011), are standard approaches when all individuals are genotyped. In the SNP-BLUP, already proposed by Meuwissen et al. (2001), the phenotypes in the vector **y** are modelled as a function of all individual SNP effects β_j, which have the same variance σ^2_β:

$$\boldsymbol{y} = \mathbf{1}\mu + \sum_{j=1}^{M} \boldsymbol{x}_j \beta_j + \boldsymbol{e},$$

where $\boldsymbol{x}_j$ is the vector of centred and scaled genotypes for marker *j*. The model can also be applied using centred genotypes $\boldsymbol{z}_j$. The SNP effects are normally distributed as $\beta_j \sim N(0, \sigma^2_\beta)$. The genomic estimated breeding values (GEBV) correspond to the polygenic effects stored in **g** and can be estimated as:

$$\boldsymbol{g} = \sum_{j=1}^{M} \boldsymbol{x}_j \beta_j.$$

In this model, the variance of the phenotypes is:

$$\text{var}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{X}'\sigma_\beta^2 + \boldsymbol{I}\sigma_e^2.$$

The SNP effect variance is a function of the total genetic variance σ_g^2, where $\sigma_\beta^2 = \sigma_g^2/M$. This model, sometimes called ridge regression, is equivalent to the GBLUP presented earlier, where phenotypes are modelled directly as a function of individual polygenic terms g_i stored in $\boldsymbol{g}$:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{g} + \boldsymbol{e}.$$

These polygenic terms are normally distributed $\boldsymbol{g} \sim (0, \boldsymbol{G}\sigma_g^2)$, where $\boldsymbol{G}$ is the GRM (see previous section for more details).

Note that when $\boldsymbol{A}$, the pedigree relationship matrix, is used instead of $\boldsymbol{G}$, the model corresponds to the 'traditional' selection used before GS. In animal breeding, the GRM computed using centred genotypes ("VanRaden1") is most commonly used (VanRaden, 2008), whereas in humans, GRM are more often computed using scaled and centred genotypes ("Yang" or "VanRaden2") (Yang et al., 2011a). Other rules for constructing GRMs have been proposed (e.g. Speed et al., 2012), where markers are "optimally" weighted according to their MAF or LD scores.

In animal breeding, extensions of GBLUP and SNP-BLUP have been developed to use information from ungenotyped animals with phenotypes (Christensen and Lund, 2010; Legarra et al., 2009). These represent often the largest fraction of the population. These extensions called single step GBLUP (ssGBLUP) (Christensen and Lund, 2010; Legarra et al., 2009) or SNP-BLUP combine pedigree and genomic BLUP. Integration of these ungenotyped individuals results in higher prediction accuracy and can also correct for some biases. In addition, this procedure has been shown to be more efficient that multistep procedures where GBLUP is first run and information from pedigree BLUP are integrated in a second step (Bradford et al., 2019). Therefore, ssGBLUP is the approach used for genomic evaluations in many countries.

In addition to GBLUP, a rather large alphabet of Bayesian models has been developed for whole genome prediction. A key feature of these models is that SNP effects can have different variances (Figure 1.8), and priors are used to estimate these SNP-specific variances. In BayesA, the variance of SNP effects has an inverse chi-squared distribution $\chi^{-2}(v, S)$ (Meuwissen et al., 2001), where v and S are the number of degrees of freedom and the scale parameter, respectively. In BayesB (Meuwissen et al., 2001), only a fraction (1-π) of the SNPs have a non-zero effect. This model better reflects the sparse distribution of QTL effects and allows some variants to have a large effect. However, in both BayesA and BayesB, the posterior distribution of the SNP effect variance for each locus is heavily influenced by the prior distribution because the data provide little information per marker. Therefore, the results

are strongly influenced by the choice of prior. In addition, the proportion of SNPs with non-zero effects is typically fixed rather than learned from the data, which further affects the flexibility and adaptability of the model to traits with different genetic architecture (Gianola et al., 2009; Habier et al., 2011). To overcome these limitations, BayesCπ and BayesDπ (Habier et al., 2011) define π as an unknown parameter that must be inferred from the data. In addition, effect sizes have the same variance in BayesCπ, making it less sensitive to the choice of priors, whereas each locus still has its own variance in BayesDπ. To capture different effect sizes of causal variants without increasing computational complexity, BayesR (Erbe et al., 2012) defines marker effects β_j as a mixture of four normal distributions with different variances:

$$p(\beta_j|\pi,\sigma_g^2)\sim\pi_1 N(0,0*\sigma_g^2)+\pi_2 N(0,10^{-4}*\sigma_g^2)+\pi_3 N(0,10^{-3}\sigma_g^2)+\pi_4 N(0,10^{-2}\sigma_g^2).$$

The model includes a null effect distribution and markers can contribute up to 1% of the genetic variance although different models with different numbers of distributions and effect sizes can be defined (Erbe et al., 2012; Moser et al., 2015). The proportions of variants within the four categories are estimated from the data. Although the same distribution is used for all markers to improve computational efficiency, the use of a mixture of four distributions has been shown to improve performance in both prediction and mapping (Moser et al., 2015). Finally, a number of models are based on a mixture of two Gaussian distributions designed to capture small polygenic effects (associated with most markers) and large marker effects (for only a few loci). This is the case for example of the Bayesian Sparse Linear Mixed Model or BSLMM (Zhou et al., 2013), Bolt-LMM (Loh et al., 2015b), BayesGC (Meuwissen et al., 2021). Although the models have common features, their implementation may differ. For example, Bolt-LMM tests a predefined set of 18 parameter combinations to increase computational efficiency. Other models have been proposed but are not presented as it is not my aim to review them all.

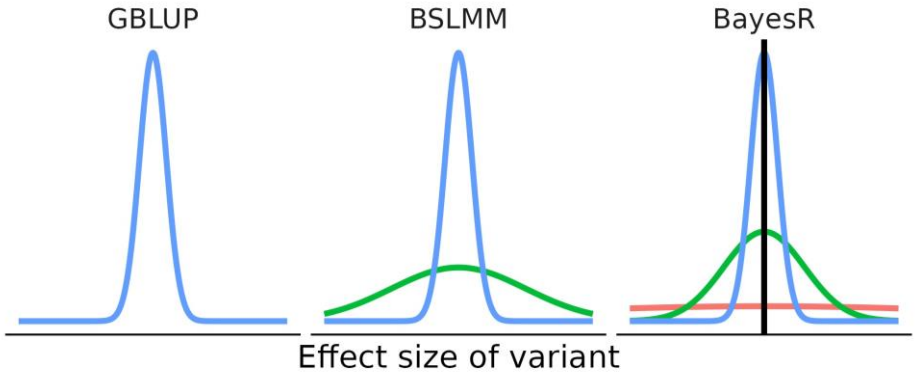


Figure 1.8. The prior distributions of SNP effect sizes are different for each model. GBLUP uses a single normal distribution, BSLMM employs a mixture of two normal distributions, and BayesR utilizes a mixture of three normal distributions and one point mass at zero.

1.4.3 Incorporation of genomic features in genomic prediction models

The original GBLUP and SNP-BLUP approaches assume the same effect size distribution for all SNP effects, although variant effect sizes may vary along the genome or across functional categories. To account for heterogeneity in SNP effect variance, several extensions of GBLUP have been proposed. In the genomic feature BLUP or GFBLUP (Edwards et al., 2016), SNPs are assigned to two groups based on genomic features or functional priors, the variants within and outside the fitted functional group. Two independent polygenic terms are then defined, one for each group:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{g}_F + \boldsymbol{g}_{\bar{F}} + \boldsymbol{e},$$

where $\boldsymbol{g}_{\mathrm{F}}$ and $\boldsymbol{g}_{\text{-F}}$ are the polygenic terms associated with the feature and the rest of the genome. These terms are normally distributed, $\boldsymbol{g}_F \sim (0, \boldsymbol{G}_F \sigma^2_{gF})$ and $\boldsymbol{g}_{\bar{F}} \sim (0, \boldsymbol{G}_{\bar{F}} \sigma^2_{g\bar{F}})$ where $\boldsymbol{G}_{\mathrm{F}}$ and $\boldsymbol{G}_{\bar{\mathrm{F}}}$ are the GRM estimated with the markers within and outside the functional category, respectively. The marker effects from the two different categories also have different variances, σ^2_{gF} and $\sigma^2_{g\bar{F}}$, which need to be estimated, e.g. by GREML.

In the MultiBLUP model (Speed and Balding, 2014), several groups of variants are fitted simultaneously, each with its own variance. The groups can be defined according to their position along the genome or according to different functional categories.

As mentioned in the section on heritability partitioning, BayesR has been extended to incorporate information from functional annotation. This was first done in BayesRC (MacLeod et al., 2016), then in BayesRCO (Mollandin et al., 2022), and finally in BayesRR-RC (Patxot et al., 2021). The main idea of these methods is that the mixture proportions are a function of the annotation group. In addition, in BayesRR-RC, each annotation group has its own variance parameter. However, both BayesRC and BayesRR-RC have the disadvantage that the SNPs are manually assigned to the different annotation groups prior to the analysis. BayesRCO is derived from BayesRC and provides two new options implemented in BayesRC+ and BayesRCπ. These options allow variants to be assigned to multiple annotation groups (Mollandin et al., 2022).

1.4.4 Application of genomic prediction models using functional information in livestock species

The ever-increasing number of whole-genome sequenced individuals, which provide reference populations for missing genotype imputation, combined with the improving functional annotation of genomes in several species, makes it possible to incorporate this functional information into genomic selection to further improve its accuracy. The use of whole genome sequence is expected to increase the accuracy of genomic selection by including causal variants in the dataset, which is particularly important

for multi-population predictions. In addition, accuracy is expected to be maintained over multiple generations as the LD between markers and causal variants does not decrease with each generation due to the recombination process. However, first attempts in livestock species did not systematically achieve higher prediction accuracy (Raymond et al., 2018; van Binsbergen et al., 2015; VanRaden et al., 2017). Some simulation studies have shown that higher accuracy is obtained when the causative variants are known, whereas using all variants from the sequence data gives at best modest improvements (Druet et al., 2014b; Meuwissen and Goddard, 2010; Pérez-Enciso et al., 2015). In addition, little gain is expected when causative variants are common and well tagged by markers from genotyping arrays. Higher gains are expected when causative variants are rare, but unfortunately these remain more difficult to impute (e.g. Druet et al., 2014b). These initial results support the use of functional annotation to improve the accuracy of genomic prediction when using sequence data.

Although genomic selection models using information from functional annotations have already been applied to livestock species, there is still a need for further investigation. Indeed, these studies have some limitations as they have rarely been carried out at the sequence level and on large cohorts. In several cases, commercial genotyping arrays have been used (Nayee et al., 2018; Rincon et al., 2011), although marker selection is then biased (i.e. not all functional variants are equally likely to be selected). These panels are enriched for common variants and capture only a small subset of causative variants, particularly when medium density arrays are used . Such marker panels may be useful for initial testing, but the use of whole genome sequence data is required to make best use of the biological information. Even when (imputed) full sequence data are available, subsets of markers have been used in a few studies (Lopez et al., 2021; VanRaden et al., 2017; Veerkamp et al., 2016) to reduce the computational burden, although this may lead to some bias and loss of causative variants, making the use of functional annotation less relevant. It is also important to remember that large datasets are required to fully exploit whole genome sequence data and functional information, but this is not always the case (Do et al., 2015). In addition, annotation information was rarely obtained from functional experiments, but rather from in silico predicted annotations, typically using VEP or similar tools, where variants are classified according to their position relative to genes (e.g. exonic, intronic, intergenic) (Lopez et al., 2021) or their predicted effect (e.g. missense variants) (MacLeod et al., 2016). Gene ontology information was also used in several studies (Edwards et al., 2016; Lingzhao et al., 2017), whereas conservation information was used less frequently (e.g. Xiang et al., 2021b, 2019a). A few studies used catalogues of eQTL (Xiang et al., 2021b, 2019a) that have strong evidence of regulatory effect, but represent only a fraction of the true eQTL, as their identification depends on the power of the study (eventually reduced by the burden of multiple testing). Ideally, these eQTL should have been identified in a large sample of the target population and in the relevant tissue. Overall, the list of genomic features included in the model remains limited compared to recent human studies based on a so-called baseline model including 24 elements (Finucane et al., 2015; Gazal et al., 2017; Speed et al., 2020).

In all these studies, the genomic feature BLUP was the most commonly used, often in combination with gene ontology information. This means that hundreds of models are tested (one per tested ontology) and that several of these may improve prediction accuracy (Edwards et al., 2016; Lingzhao et al., 2017). However, the approach does not provide a solution for combining these identified ontologies to perform prediction in a single model. In addition to GFBLUP, BayesRC has also been used in a limited number of studies, but never with the full sequence data. Another strategy commonly used to exploit biological priors is to perform a pre-selection of markers based on their annotation and then apply 'classical' models that don't exploit the biological information (MacLeod et al., 2016; Xiang, 2021). This strategy has the advantage of being more computationally efficient, but does not optimally exploit the functional information. Overall, it remains difficult to determine which strategies lead to a consistent improvement in prediction accuracy, as a variety of approaches have been applied to different traits in different populations. There is a large variation in results even for the same method, and in many cases the use of functional annotation in livestock species has not resulted in higher prediction accuracy (Abdollahi-Arpanahi et al., 2017; Xiang, 2021). Recent studies in human genetics have reported greater and more consistent benefits from using functional annotation for genomic prediction (Orliac et al., 2022; Zhang et al., 2021), suggesting that there may be further room for improvement in livestock species, particularly where more functional information is available.

Objectives

2 Objectives

Genomic selection has more than doubled genetic progress in cattle breeding over the past decade. However, the GBLUP method widely used in genomic selection still needs improvement. Improving functional annotation and integrating it into genomic selection models has the potential to significantly increase the accuracy of genomic selection. GWAS and heritability partitioning studies in human genetics are increasingly highlighting the importance of regulatory variants in complex traits. Therefore, the identification of regulatory variants in cattle holds great promise for pinpointing causative regulatory elements and improving the accuracy of genomic prediction. While significant progress has been made in human genetics with the establishment of several large consortia that are accelerating our understanding of the genetic architecture of complex traits through the identification of regulatory elements, parallel efforts in cattle are notably limited. The lack of a comprehensive regulatory map hinders our ability to unravel the genetic mechanisms controlling complex traits in cattle and to improve genomic selection methods. Therefore, this thesis has three main objectives.

The first objective of this thesis is to generate a comprehensive map of regulatory elements using a large number of samples covering almost all tissues, including those associated with economically important traits, and to use this map to identify regulatory variants. This will involve generating a catalog of regulatory regions in the bovine genome using ATAC-Seq and identifying variants within open chromatin regions. To assess the relevance of the catalog, I will then evaluate its enrichment in regulatory variants. In addition, this first study will evaluate the specificity and precision of using ATAC-Seq to identify regulatory variants, and will examine the utility of these regulatory variants for identifying causative variants.

The second objective is to investigate the genetic contribution of different functional categories of variants to the genetic variation of complex traits in Belgian Blue cattle. This will involve a comprehensive evaluation of heritability partitioning approaches in livestock populations, followed by their use to estimate the heritability explained by different functional categories for muscular development traits and height, with a focus on the contribution of regulatory elements.

The third objective is to explore strategies for using functional annotation, in particular the catalog of identified regulatory elements, to improve the accuracy of genomic selection. This will include evaluating the accuracy of genomic selection for muscular development traits in Belgian Blue cattle using whole-sequence data, evaluating the performance of genomic prediction models that use functional annotations as priors, and evaluating the accuracy of genomic selection when using subsets of variants selected based on functional annotation.

Experimental section

Experimental section

Study 1

An organism-wide ATAC-Seq peak catalogue for the bovine and its use to identify regulatory variants

Genome Research 33(10):1848-1864

Can Yuan, Lijing Tang, Thomas Lopdell, Vyacheslav A Petrov, Claire Oget-Ebrad, Gabriel Costa Monteiro Moreira, José Luis Gualdrón Duarte, Arnaud Sartelet, Zhangrui Cheng, Mazdak Salavati, D Claire Wathes, Mark A Crowe, GplusE consortium, Wouter Coppieters, Mathew Littlejohn, Carole Charlier, Tom Druet, Michel Georges & Haruko Takeda

3 Experimental section: Study 1

3.1 Summary

We herein report the generation of an organism-wide catalogue of 976,813 cis-acting regulatory elements for the bovine detected by the Assay for Transposase Accessible Chromatin using sequencing (ATAC-Seq). We regroup these regulatory elements in 16 components by non-negative matrix factorization. Correlations between the genome-wide density of peaks and transcription start sites, between peak accessibility and expression of neighboring genes, and enrichment in transcription factor binding motifs supports their regulatory potential. Using a previously established catalogue of 12,736,643 variants, we show that the proportion of single nucleotide polymorphisms mapping to ATAC-Seq peaks is higher than expected and that this is due to an ~ 1.3-fold higher mutation rate within than outside peaks. Their site frequency spectrum indicates that variants in ATAC-Seq peaks are subject to purifying selection. We generate eQTL datasets for liver and blood and show that variants that drive eQTL fall into liver and blood-specific ATAC-Seq peaks more often than expected by chance. We combine ATAC-Seq and eQTL data to estimate that the proportion of regulatory variants mapping to ATAC-Seq peaks is approximately 1 in 3, and that the proportion of variants mapping to ATAC-Seq peaks that are regulatory is approximately 1 in 25. We discuss the implication of these findings on the utility of ATAC-Seq information to improve the accuracy of genomic selection.

Resource

An organism-wide ATAC-seq peak catalog for the bovine and its use to identify regulatory variants

Can Yuan,[1] Lijing Tang,[1] Thomas Lopdell,[2] Vyacheslav A. Petrov,[1] Claire Oget-Ebrad,[1] Gabriel Costa Monteiro Moreira,[1] José Luis Gualdrón Duarte,[1] Arnaud Sartelet,[3] Zhangrui Cheng,[4] Mazdak Salavati,[4,8] D. Claire Wathes,[4] Mark A. Crowe,[5] GplusE Consortium,[5,7] Wouter Coppieters,[6] Mathew Littlejohn,[2] Carole Charlier,[1] Tom Druet,[1] Michel Georges,[1] and Haruko Takeda[1]

[1]Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège, 4000 Liège, Belgium; [2]Research and Development, Livestock Improvement Corporation, Hamilton 3240, New Zealand; [3]Clinical Department of Ruminant, University of Liège, 4000 Liège, Belgium; [4]Royal Veterinary College, Hatfield, Herts AL9 7TA, United Kingdom; [5]School of Veterinary Medicine, University College Dublin, Dublin 4, Ireland; [6]GIGA Genomics platform, GIGA Institute, University of Liège, 4000 Liège, Belgium

We report the generation of an organism-wide catalog of 976,813 *cis*-acting regulatory elements for the bovine detected by the assay for transposase accessible chromatin using sequencing (ATAC-seq). We regroup these regulatory elements in 16 components by nonnegative matrix factorization. Correlation between the genome-wide density of peaks and transcription start sites, correlation between peak accessibility and expression of neighboring genes, and enrichment in transcription factor binding motifs support their regulatory potential. Using a previously established catalog of 12,736,643 variants, we show that the proportion of single-nucleotide polymorphisms mapping to ATAC-seq peaks is higher than expected and that this is owing to an approximately 1.3-fold higher mutation rate within peaks. Their site frequency spectrum indicates that variants in ATAC-seq peaks are subject to purifying selection. We generate eQTL data sets for liver and blood and show that variants that drive eQTL fall into liver- and blood-specific ATAC-seq peaks more often than expected by chance. We combine ATAC-seq and eQTL data to estimate that the proportion of regulatory variants mapping to ATAC-seq peaks is approximately one in three and that the proportion of variants mapping to ATAC-seq peaks that are regulatory is approximately one in 25. We discuss the implication of these findings on the utility of ATAC-seq information to improve the accuracy of genomic selection.

[Supplemental material is available for this article.]

Genomic selection has had a tremendous impact on livestock breeding in the past 10 yr (e.g., García-Ruiz et al. 2016). Nevertheless, the accuracy of selection remains inferior to what may be achievable given the heritability of the selected traits. This could have a number of causes, including the size and composition of the reference population or the contribution of dominance and epistasis to the genetic architecture of the traits of interest. Another factor is that all variants are generally given an equivalent weight in computing the additive relationship between animals needed for GBLUP analyses or equivalent prior probabilities of variant effects in Bayesian approaches. Yet, only a minority of variants are causative (having a direct effect on gene function and hence phenotype), with the remainder being, at best, passenger variants in linkage disequilibrium (LD) with one or more of the causative variants. The extent of LD between causative and passenger variants is bound to be population specific, or even subpopulation specific, and is likely to fluctuate over time, and this may account in part for the observed limits in selection accuracy. It is generally believed that knowing the causative variants, or at least those that are more likely to be, may help to further improve the accuracy of genomic selection (Xiang et al. 2019).

Causative variants encompass coding and regulatory variants. Coding variants, including missense, nonsense, frameshift, splice site variants, and deletions, are easily recognized yet only account for a limited part of the genetic variance for complex phenotypes, including production traits. It is increasingly apparent that most of the genetic variation for complex traits is owing to regulatory variants that act either by perturbing the expression profile of genes located in *cis* (standard polygenic model) or, possibly, by perturbing the gene regulatory network and affecting the expression profile of a restricted number of core genes in *trans* (omnigenic model) (Liu et al. 2019). Regulatory variants are more difficult to identify as the effect of polymorphisms on the functionality of proximal and distant *cis*-acting regulatory elements remains difficult to predict. However, it is reasonable to assume that most regulatory variants are located within or in close proximity to regulatory elements, which account for an estimated ~5%–20% of genome space (Meuleman et al. 2020; The ENCODE Project

[7]A complete list of the GplusE Consortium authors appears at the end of this paper.
[8]Present address: Dairy Research and Innovation Centre, Scotland's Rural College, Barony Campus, Dumfries DG1 3NE, UK
Corresponding author: michel.georges@uliege.be
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277947.123.

33:1848–1864 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/23; www.genome.org

Consortium et al. 2020). Active regulatory elements can be recognized by virtue of evolutionary constraint (Lindblad-Toh et al. 2011) and epigenetic features, including chromatin accessibility, specific histone codes, transcriptional activity, their participation in loop structures, and transcription factor (TF) occupancy (Meuleman et al. 2020; The ENCODE Project Consortium et al. 2020).

In an effort to identify putative regulatory variants in the bovine, we herein report (1) the generation of a comprehensive catalog of bovine regulatory elements identified using assay for transposase accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al. 2013) in 63 tissue types; (2) the generation of a catalog of common bovine variants that map to identified proximal and distal regulatory elements; (3) the demonstration that variants driving expression quantitative trait loci (eQTL) in liver and blood are more likely to map to regulatory elements that are active in the cognate tissues and, hence, that variants in these regulatory elements are more likely to be causative; (4) estimates of the proportion of regulatory variants that map to ATAC-seq peaks as well as the proportion of variants mapping to ATAC-seq peaks that are regulatory; and (5) a retrospective evaluation of the utility of this catalog for the identification of regulatory variants known in livestock species.

Results

Generating a catalog of bovine *cis*-acting gene regulatory elements

To generate a bovine catalog of open chromatin regions using ATAC-seq, we collected 106 samples corresponding to 68 tissue types (Fig. 1A; Supplemental Tables S1, S2). Most samples (73%) were collected from the same juvenile Holstein male. The remainder (27% including gonads and mammary gland) were collected from nine additional animals (Supplemental Tables S1, S2). Fresh and frozen samples were subjected to ATAC-seq using standard procedures with two concentrations of tagmentation enzyme (Buenrostro et al. 2013; Corces et al. 2017). We sequenced a total of 185 libraries to an average of 31.8 million paired-end reads per library (Supplemental Table S3). To these in-house-generated data, we added publicly available ATAC-seq data (15 data sets) from five additional tissues/cell types (Fang et al. 2019; Halstead et al. 2020a,b; Johnston et al. 2021). Reads were mapped to the bovine genome (ARS-UCD1.2) with Bowtie 2 (Langmead and Salzberg 2012) and ATAC-seq peaks called with MACS2 (Zhang et al. 2008) following ENCODE's recommendations (https://www.encodeproject.org/atac-seq/). Data sets passing quality control (89/106 in-house-generated data, i.e., 84%) and corresponding to technical replicates were merged (per biosample), resulting in a total of 104 ATAC-seq data sets (89 in-house and 15 public) representing 63 tissue types (58 in-house and five public). Pearson's correlations between technical and biological replicates (normalized read counts across 500-bp windows covering the entire genome) exceeded 0.89 and 0.85, respectively (Supplemental Figs. S1, S2).

MACS2 yielded an average of 76,919 peaks per sample in ATAC-seq mode (range: 15,420–238,210 peaks) and 51,838 peaks per sample in ChIP-seq mode (range: 14,594–201,757 peaks) (Supplemental Fig. S3; Supplemental Table S4). We merged ATAC-seq-mode and ChIP-seq-mode peaks separately across 104 samples following the method of Meuleman et al. (2020), and joined the resulting peaks (when overlapping). This yielded a total of 976,813 reference peaks (excluding the Y Chromosome and unanchored scaffolds) with core and consensus segments (empirical confidence bounds of aggregates of peak summits and regions, respectively) (Supplemental File S1; Meuleman et al. 2020). Core and consensus segments amounted, respectively, to 134 Mb and 264 Mb, or 5.1% and 10.0% of genome space. Of these, 41,841 peaks (4.3%) were located in promotor regions (defined as 1 kb upstream of to 0.1 kb downstream from the transcription start sites [TSSs]) of 33,579 Ensembl reference transcripts (out of 43,512) and were referred to as "proximal," whereas the remaining 934,972 peaks were considered "distal." The median consensus size of the proximal peaks (306 bp) was larger than that of the distal peaks (216 bp; $P_{\text{Wilcoxon}} < 2.2 \times 10^{-16}$) (Fig. 1B). The proximal peaks were "open" in more tissues than the distal peaks (i.e., distal peaks were more often tissue-specific; $P_{\text{Wilcoxon}} < 2.2 \times 10^{-16}$) (Fig. 1C). The accessibility of "open" peaks was higher for proximal (14.0-fold increase of read depth over background) than for distal peaks (7.3-fold increase of read depth over background; $P_{\text{Wilcoxon}} = 4.3 \times 10^{-19}$) (Fig. 1D). Of note, the distribution of genomic evolutionary rate profiling (GERP) scores (Cooper et al. 2005; Davydov et al. 2010) was overdispersed for both proximal and distal peaks, showing an excess of positions with higher and lower substitution rates than expected (under neutrality) compared with flanking regions (Fig. 1E).

ATAC-seq peaks were unevenly distributed across the genome, both between and within chromosomes (Supplemental Fig. S4A,B). The density of ATAC-seq peaks was highest for Chromosome 19 and lowest for Chromosomes 6 and 12. Chromosome X was also particularly poor in ATAC-seq peaks, but this could be owing to its hemizygosity in a majority of male samples. The density of ATAC-seq peaks was highly correlated with the density of TSSs ($r = 0.52$) (Supplemental Fig. S4C).

We used unsupervised nonnegative matrix factorization (NMF) according to the method of Meuleman et al. (2020) to decompose the 976,813-peak × 104-sample matrix in 16 components (Fig. 1F; Supplemental Tables S5, S6; Supplemental File S1; Supplemental Figs. S5, S6). NMF converts each sample and each peak into a linear combination of these 16 components, that is, a weighted sum of the 16 components. Twelve of the 16 components could be readily assigned to recognizable bodily systems as they would be dominant (>30% of the weight) in anatomically related samples. They were labeled accordingly: central nervous system (CNS), cerebellum, immune system, digestive tract, ruminal epithelium, lower respiratory, upper respiratory, muscle, liver, endocrine, mammary gland, and testis. Accordingly, these 12 components dominated about 629,870 ATAC-seq peaks (64.5%) characterized by tissue-specific accessibility. Three components corresponded, respectively, to eight-cell embryo, morula, and inner cell mass (ICM) and dominated a very distinct set of about 213,305 ATAC-seq peaks (21.8%), of which 54,498 (5.6%), 76,175 (7.8%), and 14,108 (1.4%) were eight-cell, morula, and ICM specific, respectively. A set of 26,414 (2.7%) peaks was shared by multiple, yet at first glance, anatomically unrelated samples. The meaning of this 16th component, whether biological or technical, remains unclear. It is referred to as "undefined." Of note, the 16th component dominated the sample types that were hard to dissociate. Finally, one group corresponded to about 107,224 (11.0%) peaks that were shared by the majority of samples and characterized by uniform weights for the 16 components (≤30% for any component). They are referred to as "ubiquitous" peaks and account for 59.6% of proximal peaks assigned to housekeeping genes (Supplemental Table S6).

NMF decomposition uses a binary matrix summarizing the presence (1) versus absence (0) of the 976,813 peaks in the 104

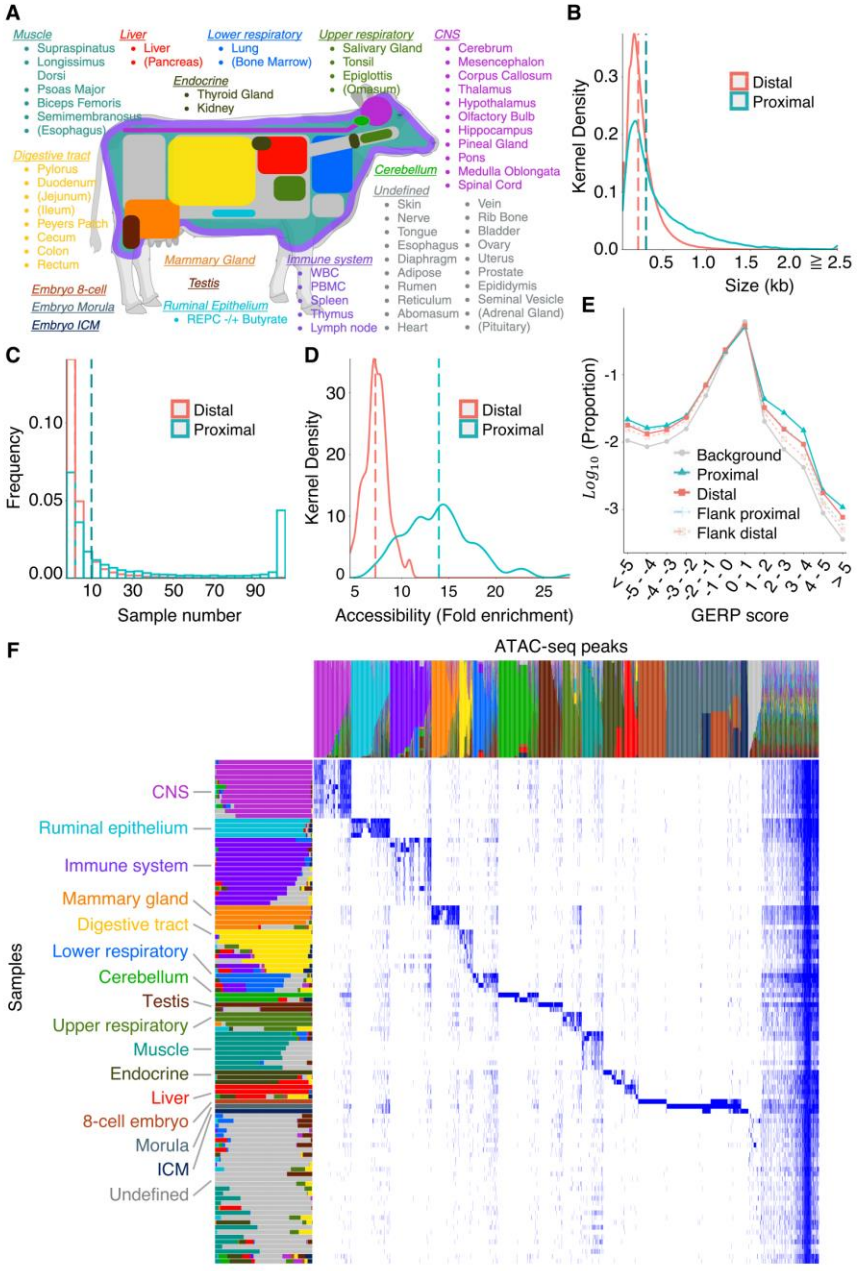


Figure 1. Generation of an organism-wide catalog of *cis*-acting regulatory elements for the bovine. (*A*) Sixty-three tissue types with ATAC-seq data analyzed in this work. Novel ATAC-seq data were generated for 58 tissue types (89 samples), and public ATAC-seq data were downloaded for five (15 samples). Tissue types are grouped and colored based on the nonnegative matrix factorization (NMF) analysis described in *D*. Tissues are parenthesized when the largest NMF component in the tissue explains <50% of the total weight. This figure was created with BioRender (https://www.biorender.com). (*B*) Size distribution of proximal (green) and distal (red) ATAC-seq peaks (consensus peaks). (*C*) Distribution of the number of samples in which proximal (green) and distal (red) ATAC-seq peaks are open. (*D*) Distribution of the accessibility (fold-increase in coverage over background) of proximal (green) and distal (red) ATAC-seq peaks. The vertical dotted lines in *B*, *C*, and *D* correspond to the medians. (*E*) Distribution of GERP scores for nucleotide positions within proximal (solid green) and distal (solid red) ATAC-seq peaks, within sequence segments of same size immediately flanking proximal (dotted green) and distal (dotted red) ATAC-seq peaks, and across the entire genome (gray). The proportion of nucleotide positions without GERP score is not shown. (*F*) Decomposition of the 976,813-peak × 104-sample matrix in 16 components by nonnegative matrix factorization (NMF) following the method of Meuleman et al. (2020). As a result, each peak and each tissue sample are represented as a linear combination of the 16 components, which are color-coded in the graph. The lengths/heights of the bars measure the loading factor of the corresponding component for each of the tissue samples/peaks. Anatomically related samples typically have the same dominant component and have been ordered accordingly (Supplemental Table S5). The peaks that are predominantly active in the cognate tissue samples are dominated by the same component and are ordered accordingly. Thirty-one samples did not show clear tissue-specific peaks; their ATAC-seq profiles were dominated by the "ubiquitous" peaks shared by nearly all samples and, to a lesser extent, by a group of peaks assigned to the 16th "undefined" NMF component (shown in gray).

samples. We also performed hierarchical clustering of the samples (Ward D2 method) (R Core Team 2023) based on a quantitative measure of the accessibility of distal peaks. This approach grouped the samples largely by the NMF component (Supplemental Fig. S7). Of note, the cerebellar samples (assigned to the NMF07_Cerebellum group) formed a distinct cluster yet were closest to the remaining CNS samples (NMF01_CNS group). Also, the ruminal epithelial primary cells (NMF02_Ruminal epithelium) formed a distinct cluster, with embryonic samples (rather than digestive tract samples) as sister clade. This suggests that culturing these cells profoundly affects the epigenetic profile of these cells, apparently toward a proliferative stem-cell like phenotype.

We evaluated the added value of analyzing extra samples, first in terms of discovery of new ATAC-seq peaks. To that end, we ranked samples by the decreasing number of newly uncovered ATAC-seq peaks (Supplemental Fig. S8A). When limiting ourselves to the 97 postnatal tissue samples, the number of newly discovered peaks saturated at about 725,000 after approximately 75 samples, suggesting that our library of ATAC-seq peaks includes the majority of regulatory elements accessible after birth. However, adding only three embryonic samples (and, to a lesser extent, primary cultured cells) uncovered an extra tier of around 200,000 peaks. This suggests that substantially more developmental stage–specific regulatory elements remain to be uncovered and that the analysis of additional fetal samples, for instance, is warranted. An additional value in analyzing more tissue types is to determine in which tissue types known regulatory elements are accessible and in which tissues they are not (Supplemental Fig. S6). The majority of peaks uncovered in a given sample are neither unique for the sample nor shared with all others but rather shared with a variable number of other samples (not necessarily from the same NMF component) which are, hence, not obvious to predict (Supplemental Fig. S8B, C). Finally, we examined the relative merits of analyzing more sample types with ATAC-seq only versus fewer sample types with multiple assays. To that end, we evaluated the overlap between the peaks identified by Kern et al. (2021) in eight tissue types (adipose, cerebellum, brain cortex, hypothalamus, liver, lung, skeletal muscle, spleen) using ATAC-seq combined with ChIP-seq (H3K4me1, H3K4me3, H3K27ac, CTCF) with our own catalog. To make for a better comparison, we reanalyzed Kern's ATAC-seq data in "narrow-peak" mode (as opposed to the "broad-peak" mode used by Kern et al.). The vast majority (93.5%) of Kern's ATAC-seq peaks overlapped with ours, as expected. Kern's ATAC-seq peaks overlapped with 69% of their H3K4me3 peaks (i.e., active promotors), 43% of their H3K4me1 peaks (i.e., active enhancers), 44% of their H3K27ac peaks (i.e., active promotors and enhancers), and 42% of their CTCF peaks. Similarly, a subset of our ATAC-seq data from the corresponding eight tissues overlapped with 76% of H3K4me3 peaks, 50% of H3K4me1 peaks, 49% of H3K27ac peaks, and 48% of CTCF peaks. In comparison, our complete ATAC-seq peak catalog overlapped with 89% of H3K4me3 peaks, 73% of H3K4me1 peaks, 71% of H3K27ac peaks, and 69% of CTCF peaks (Supplemental Fig. S8D). Thus, it appears that analyzing more sample types by ATAC-seq compensates to some extent for the use of a single assay as it recovers regulatory elements that are missed if performing only ATAC-seq on fewer sample types. This finding also suggests that the same regulatory element may adopt distinct epigenetic configurations, presumably associated with distinct functional states, captured by distinct assays in different tissues.

We searched for TF binding motifs enriched in tissue-specific and ubiquitous ATAC-seq peaks using HOMER (Fig. 2A; Supplemental Tables S7, S8; Heinz et al. 2010). For each component, binding motifs were found to be very significantly enriched, in good agreement with previous reports for bovine tissue-specific *cis*-acting regulatory elements and/or tissue-specific function of the corresponding TF in other species. Moreover, using publicly available RNA-seq information, we found that 35 of the cognate TFs were more highly expressed in the corresponding tissues compared with all other ones (Supplemental Table S7).

We matched 91 of our tissue-specific ATAC-seq data with publicly available RNA-seq data from 56 bovine tissues (Supplemental Table S9), and computed correlations between gene expression and accessibility of ATAC-seq peaks mapping within 1 Mb from the gene's TSS (Fig. 2B). Correlations were overdispersed, showing too many positive but also negative correlations. Indeed, any peak that is specific for a given tissue type will be positively correlated with any gene that is specifically expressed in that same tissue type. These correlations are therefore not indicative of *cis* interactions between peaks and their target gene(s). However, positive correlations were increasing in numbers (and becoming more significant) as the distance between peak and gene decreased. This inflation of positive correlations over the background (i.e., at distances ≥ 750 kb) was highly significant for gene-peak distances up to ~250 kb. This supports the common occurrence of direct *cis* interactions between enhancer peaks and target genes, at least up to such distances. The effect was slightly more pronounced for peaks located downstream from the TSSs than peaks located upstream of the TSSs. The same trend was not observed when repeating the same analyses with negative correlations (Fig. 2B). In fact, we observed a slight deflation of negative correlations (becoming less negative and less significant) as the distance between the peak and gene decreased below ~40 kb. This suggests that few ATAC-seq peaks act as *cis* silencers on target genes.

Generating a catalog of common variants mapping to *cis*-acting regulatory elements

We used a previously established catalog of 11,030,905 single-nucleotide variants (SNVs) and 1,705,738 short (≤265-bp) insertion–deletion variants (indels) obtained by analyzing 264 Holstein-Friesian (HF) whole-genome sequences (average, 25.2-fold depth; range, 15.2 to 47.1) (Oget-Ebrad et al. 2022) using GATK (Poplin et al. 2018). Of these, 1,256,997 SNVs (11.4%) and 133,394 indels (7.8%) mapped to ATAC-seq peaks (Supplemental File S2).

We studied the proportion of indels falling within versus outside ATAC-seq peaks separately for the following genome compartments: TSSs, 100 bp upstream of to 1 kb downstream from transcription termination sites (TTSs), exons, introns, and intergenic regions (Fig. 3B). The proportion of indels mapping to ATAC-seq peaks was significantly below expectations for TSSs (i.e., proximal peaks) ($P = 1.4 \times 10^{-33}$), TTSs ($P = 1.2 \times 10^{-52}$), introns ($P < 1.0 \times 10^{-100}$), and intergenic regions ($P < 1.0 \times 10^{-100}$). These effects were even stronger when considering common indels only (minor allele frequency [MAF] > 0.05). This is the expected signature of purifying selection acting on functionally important elements. Of note, the proportion of all indels (but not common indels) was slightly higher than expected (p = 0.18) for the exonic compartment. This effect became significant ($P = 1.8 \times 10^{-10}$) when restricting the analysis to open reading frames (ORFs) (i.e., ignoring 5′ and 3′ untranslated regions).

In contrast, the proportion of SNVs mapping to ATAC-seq peaks was significantly higher than expected for all five genomic compartments: TSSs ($P = 2.7 \times 10^{-21}$), TTSs ($P = 3.9 \times 10^{-19}$), exons

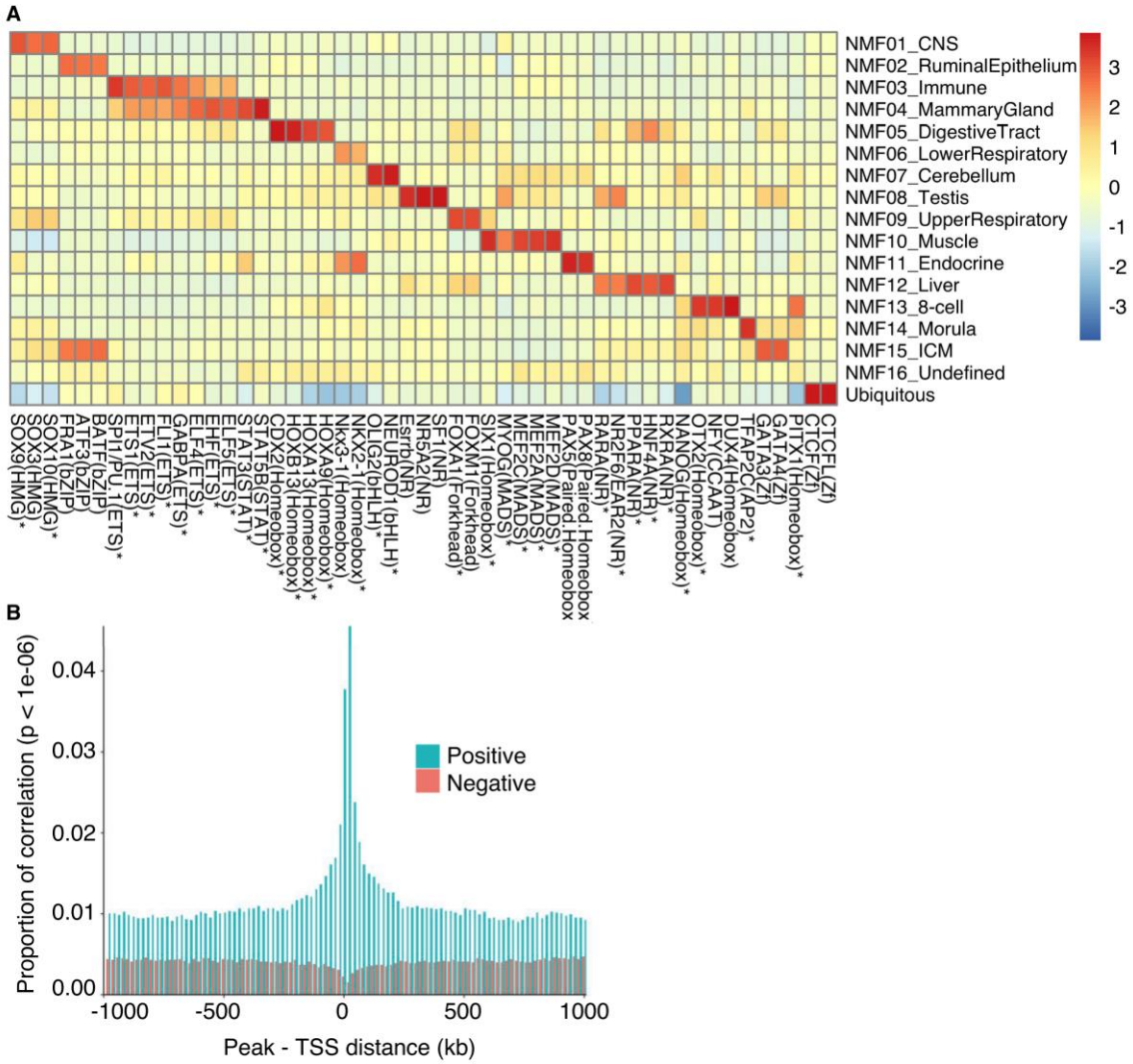


Figure 2. Open chromatin regions are enriched in *cis* regulatory elements. (*A*) TFs (*x*-axis) whose binding motifs are enriched in tissue type–specific ATAC-seq peaks assigned to the corresponding NMF components (*y*-axis). The color code measures the excess in the percentage of peaks encompassing the corresponding motif over background, scaled (*Z*-score) across NMF components. TFs that are also more strongly expressed in tissues corresponding to that component compared with other tissues (Supplemental Table S7) are marked by asterisks. (*B*) Proportion of significant, across tissue type, correlations ($P<10^{-6}$) between ATAC-seq peak accessibility and gene expression as a function of the distance between the TSS and the peak. Green indicates positive correlations; red, negative correlations.

($P=4.4\times10^{-41}$), introns ($P<1.0\times10^{-100}$), and intergenic regions ($P<1.0\times10^{-100}$) (Fig. 3A). The effect was reduced when considering common SNVs only but was still significant for TTSs ($P=2.6\times10^{-3}$), exons ($P=8.7\times10^{-6}$), introns ($P<10^{-100}$), and intergenic regions ($P<10^{-100}$). This is counter-intuitive as ATAC-seq peaks are assumed to be functionally important elements and, hence, subject to purifying selection that should result in fewer than expected number of variants. Only for TSSs were common SNVs significantly underrepresented in ATAC-seq peaks ($P=8.2\times10^{-11}$). These observations corroborate recent findings in *Arabidopsis thaliana* (Monroe et al. 2022) and humans (Kaiser et al. 2021; Luquette et al. 2022). They may be related to the reduced efficiency of RNase H2–dependent repair of erroneously incorporated nucleotides during pol α-dependent initiation of DNA replication of Okazaki fragments (Reijns et al. 2015), or of nucleotide excision repair (Sabarinathan et al. 2016), at sites where proteins, including TFs, bind DNA. In agreement with this hypothesis, the density of singletons (supposed to be enriched in recent mutations and hence used as surrogate for de novo mutations [DNMs]) was higher in ATAC-seq peaks than in flanking sequences (Fig. 3C,D). Knowing that the expected number of singletons per base pair equals $4N\mu$ independently of sample size (Nielsen and Slatkin 2013), and under

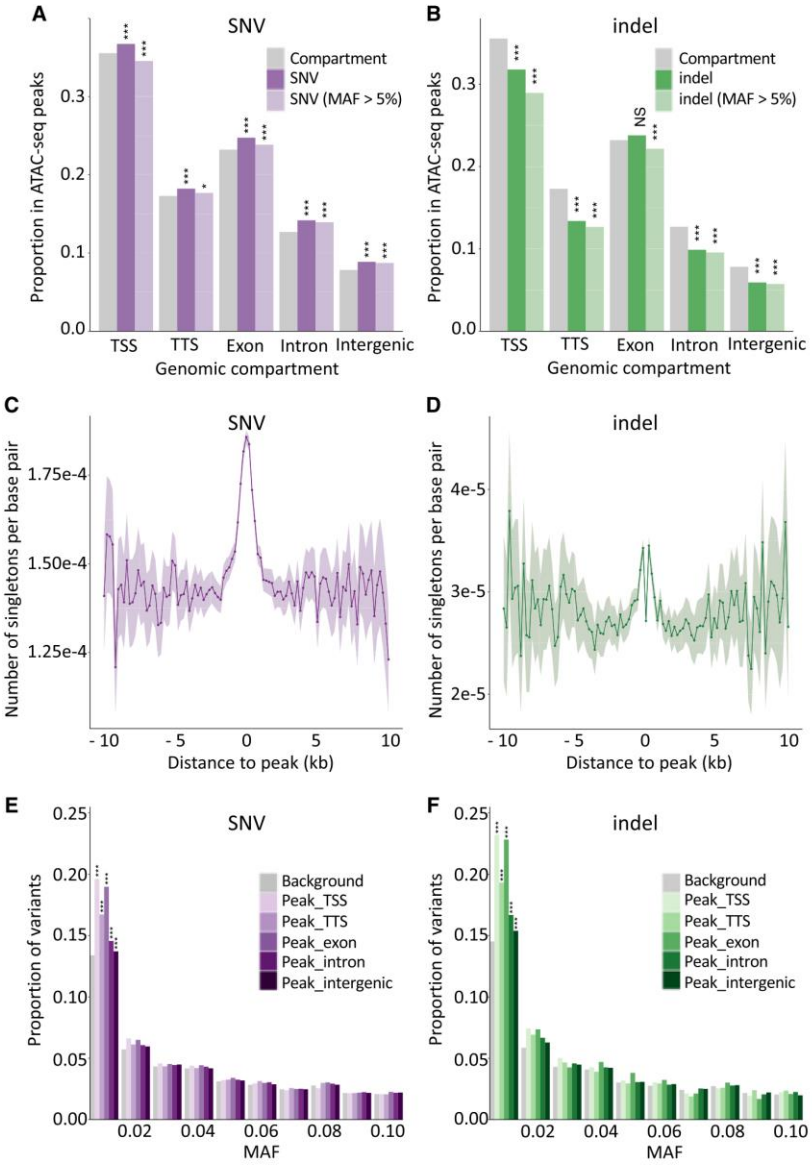


Figure 3. Open chromatin regions are mutational hotspots yet are subject to purifying selection. (*A*) Proportion of SNVs that map in ATAC-seq peaks for different genome compartments (*x*-axis: TSS, TTS, exon, intron, intergenic). Gray indicates the proportion of a corresponding genome compartment that is occupied by ATAC-seq peaks; dark purple, all SNVs; and light purple, common SNVs (MAF > 0.05). (***) $P \leq 0.001$, (*) $P \leq 0.05$. (*B*) As in *A* for indels. Gray indicates the proportion of a corresponding genome compartment that is occupied by ATAC-seq peaks; dark green, all indels; and light green, common indels (MAF > 0.05). (***) $P \leq 0.001$, (NS) nonsignificant. (*C*) Number of singleton SNVs (per interrogated base pair) in 264 whole-genome-sequenced Holstein-Friesian (HF) animals in nonoverlapping 200-bp windows at increasing distances from the center of ATAC-seq peaks. The shaded area corresponds to 2 × SD for the corresponding window. Fluctuation increases with distance as the number of windows decreases. (*D*) As in *C* for singleton indels. The excess near the ATAC-seq peak centers is clearly visible despite the drop at their very center (assumed to reflect purifying selection). (*E*) Folded SFS ($0.0 < \text{MAF} \leq 0.1$) for SNVs mapping within ATAC-seq peaks assigned to different genome compartments (purple range indicates TSS, TTS, exon, intron, intergenic) compared with SNVs outside peaks. (***) $P \leq 0.001$. (*F*) Folded SFS ($0.0 < \text{MAF} \leq 0.1$) for indels mapping within ATAC-seq peaks assigned to different genome compartments (green range indicates TSS, TTS, exon, intron, intergenic) compared with indels outside peaks. (***) $P \leq 0.001$.

some simplifying assumptions, the SNV mutation rate may be about 1.3 times higher within than outside ATAC-seq peaks.

In melanoma, the rate of somatic mutations is increased about fivefold at accessible TF binding sites, and this is thought to be because of hampered nucleotide excision repair by bound TFs (Sabarinathan et al. 2016). To verify whether the excess of SNVs in our ATAC-seq peaks was likewise concentrated in TF binding motifs, we identified 386,812 NMF component-specific peaks

(weight of one component >90%) and, within those, the positions of the 10 most enriched binding motifs for that component (de novo enrichment analysis) (Supplemental Table S8). We then checked whether SNVs mapping to the corresponding peaks would fall more often within than outside of the binding motifs. There was no evidence for a preferential location of SNVs in binding motifs, whether at the motif, NMF component, or global level. If any trend, the proportion of SNVs in binding motifs was slightly inferior to their corresponding peak occupancy (global $P = 0.06$) (Supplemental Table S10).

To further check whether variants mapping to ATAC-seq peaks, including SNVs, might be under purifying selection as expected, we compared the folded site frequency spectrum (SFS) of variants mapping within ATAC-seq peaks for the five genomic compartments in the 264 sequenced animals, with the folded SFS of all variants flanking peaks. The proportion with MAF ≤ 0.01 was higher for variants mapping in ATAC-seq peaks, and this applied both to indels and to SNVs. The effect was strongest for TSSs and exons (Fig. 3E,F).

Taken together, our data support the notion that ATAC-seq peaks are mutational hotspots, explaining the observed excess of SNVs, yet are subject to enhanced purifying selection, accounting for the depletion in indels and the shift of the SFS toward low frequencies for both SNVs and indels. This hypothesis may also account for the overdispersed GERP scores (Fig. 1E).

Of the 1,390,391 genetic variants mapping to ATAC-seq peaks, 847,831 SNVs and 86,673 indels are common with MAF > 0.05 in the sequenced animals (Supplemental File S2). These are prime candidates to receive particular attention when computing genomic breeding values in genomic selection.

Identifying bovine *cis* eQTL in liver and blood

To evaluate whether our catalog of "ATAC-seq variants" is enriched in regulatory variants, we performed eQTL analyses. We collected whole-blood and liver samples from, respectively, 224 and 176 HF cows and performed RNA-seq using standard procedures (Supplemental Table S11; Lee et al. 2021; Wathes 2021a,b). The reads were mapped to the bovine genome (ARS-UCD1.2) using HISAT2 (Kim et al. 2015) and read coverage for 27,233 reference genes (bosTau9.ensGene.gtf, v101) estimated using StringTie (Pertea et al. 2015). Gene count data were normalized within sample using DESeq2 (Love et al. 2014) following the method of Anders and Huber (2010) and across samples using inverse normal transformation. After filtering out lowly expressed genes, 14,289 genes were retained for eQTL analyses in blood and 15,458 in liver. All samples were genotyped with a high-density SNV array interrogating 777,962 variants and imputed to whole genome using Minimac4 (Das et al. 2016) and the 264 sequenced Holstein animals as reference. This yielded usable genotypes for 8.4 million SNVs and 1.3 million indels with MAF > 0.02. *Cis* eQTL analyses (variants within 1 Mb from gene's TSS) were conducted using residuals corrected for hidden PEER factors (Stegle et al. 2010), country (of origin of the samples), and polygenic effects estimated with GenABEL (Aulchenko et al. 2007), under an additive model using QTLtools (Delaneau et al. 2017). Nominal *P*-values were corrected for multiple testing within the 2-Mb *cis* window by permutation. The best-corrected *P*-value was retained for each gene and converted to FDR value by tissue type. *Cis* eQTLs with FDR < 0.05 were considered significant.

We obtained 7817 significant *cis* eQTLs in blood and 6172 in liver (Supplemental Table S12). These numbers correspond to 39.9% and 54.7% of interrogated genes, respectively, and are comparable to findings in humans (https://gtexportal.org/home/tissueSummaryPage). Leading variants tended to concentrate (and −log(p) values hence to be highest) in the vicinity of the TSSs (Supplemental Fig. S9). The proportion of significant blood eQTLs that would also operate in liver was estimated at 67% using π_1 following the method of Storey and Tibshirani (2003), whereas the proportion of significant liver eQTLs that would also operate in blood was estimated at 78%.

We defined "credible variant sets" (i.e., sets of variants that are more likely to include the causative variants that are functionally driving the observed *cis* eQTL effect) as the leading variant plus the variants in LD with it at threshold r^2-value of 0.9. The median size of credible sets was 12, ranging from one to 2870.

Variants driving eQTL are preferentially mapping in ATAC-seq peaks

If variants mapping to ATAC-seq peaks are indeed enriched in causative variants, they should be enriched in the credible sets driving *cis* eQTL effects. The significance of the overlap between *cis* eQTL credible sets and ATAC-seq peaks was evaluated by permutation following the method of Trynka et al. (2015) (Fig. 4A,B; Supplemental Table S13). Analyses were conducted by NMF component (assigning peaks to their dominant component). Credible sets for blood-specific *cis* eQTLs were most significantly ($P \leq 0.0001$) enriched in variants mapping to ATAC-seq peaks assigned to the immune and ubiquitous NMF components. Credible sets for liver-specific *cis* eQTLs were most significantly ($P \leq 0.0001$) enriched in variants mapping to ATAC-seq peaks assigned to the liver and ubiquitous NMF components. The enrichment in tissue-specific ATAC-seq peaks (immune for blood eQTL, and liver for liver eQTL) was driven by distal peaks, whereas the enrichment in ubiquitous ATAC-seq peaks was driven by both proximal and distal peaks (Fig. 4A,B; Supplemental Table S13).

Estimating the proportion of regulatory variants mapping in ATAC-seq peaks and the proportion of variants mapping in ATAC-seq peaks that are regulatory

The utility of ATAC-seq data for the identification of regulatory variants underpinning the heritability of complex traits depends on the proportion of regulatory variants that map to ATAC-seq peaks (i.e., the sensitivity or ratio of true positives/[true positives + false negatives]), and the proportion of regulatory variants among variants mapping to ATAC-seq peaks (i.e., the precision or ratio of true positives/[true positives + false positives]). The combination of ATAC-seq and *cis* eQTL information provides an opportunity to estimate these parameters. For example, if all *cis* eQTLs are driven by a regulatory variant mapping to an ATAC-seq peak, all credible sets should contain at least one variant mapping to an ATAC-seq peak. We developed a maximum likelihood-based approach (see Methods) to estimate the proportion of *cis* eQTLs driven by regulatory variants in ATAC-seq peaks from the observed excess of credible set variants mapping in ATAC-seq peaks (over the proportion of the genome occupied by ATAC-seq peaks). The parameters estimated by this approach were then used to estimate the proportion of regulatory variants among those that map to ATAC-seq peaks (see Methods).

We applied this approach to the 7817 blood and 6172 liver eQTLs. It yielded estimates of 0.34 (blood) and 0.32 (liver) for sensitivity, and 0.044 (blood) and 0.041 (liver) for precision. In other words, approximately one out of three regulatory variants maps to

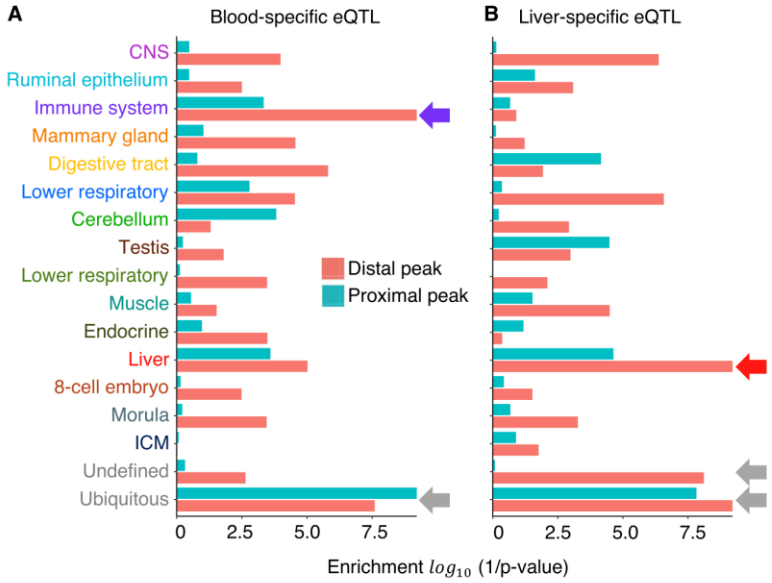


Figure 4. Open chromatin regions are enriched in *cis*-regulatory variants. Enrichment of variants mapping to NMF component-specific ATAC-seq peaks in credible sets ($r^2 \geq 0.9$ with the lead variant) of 3857 blood-specific and 2212 liver-specific *cis* eQTL, evaluated by following the method of Trynka et al. (2015). The *x*-axis shows statistical significance (log(1/p)) of the enrichment; *y*-axis, NMF component. Green indicates proximal peaks; red, distal peaks. (*A*) Blood-specific eQTLs. (*B*) Liver-specific eQTLs.

an ATAC-seq peak, and approximately one in 25 variants mapping to ATAC-seq peaks is regulatory.

Retrospective evaluation of the utility of ATAC-seq information for the identification of known regulatory variants in livestock

Prior positional cloning studies conducted in livestock identified at least three regulatory variants influencing economically important quantitative traits. The first is the *IGF2*-intron3-3072 variant in the pig that precludes binding of the ZBED6 repressor to a conserved silencer element in intron 3 of the *IGF2* gene, leading to illegitimate postnatal expression of the paternal *IGF2* allele in striated muscle and, hence, muscular hypertrophy (Van Laere et al. 2003; Markljung et al. 2009). The sequence conservation of the corresponding silencer element suggests that it operates in a similar manner across species. Nevertheless, there was no evidence in our ATAC-seq peak catalog of any peak overlapping the orthologous position of the quantitative trait nucleotide (QTN; bosTau9 Chr 29: 49,408,409), whether tissue specific (including in muscle) or tissue shared (Fig. 5A). The second is the ovine rs10721113 callipyge QTN that perturbs the function of a putative silencer element highly conserved among placental mammals, located in the *GTL2-DLK1* intergenic region, that—in wild-type sheep—suppresses postnatal muscular expression of a cluster of imprinted genes (including the paternally expressed *DLK1* and *PEG11* genes). Animals inheriting this mutation from their sire express the callipyge muscular hypertrophy (Freking et al. 2002; Georges et al. 2004). There was no clear evidence of a peak overlapping the orthologous position of the QTN (bosTau9 Chr 21: 65,691,395) in postnatal skeletal muscle ATAC-seq peaks, as would be expected. There was such a peak in testes and to a lesser extend in tongue, but—in hindsight—this would not have been considered strong support for the causality of the corresponding variant, and the significance of this finding—if any—remains unclear (Fig. 5B). The third example concerns a credible set of eight noncoding variants affecting bovine stature and several other traits by perturbing the expression of *PLAG1* and possibly other genes in its vicinity (Karim et al. 2011). Of the eight variants, only the two that map to the supposedly bidirectional promotor between *PLAG1* and *CHCHD7* (rs20982 1678: (CCG)9/(CCG)11 microsatellite and rs210030313: A/G SNV) overlap with strong ubiquitous ATAC-seq peaks (Fig. 5C). Of note, previously conducted reporter and EMSA assays supported the causality of both variants (Karim et al. 2011). Further supporting their causality, the ATAC-seq data reveal an allelic imbalance for the rs210030313 SNV (Fig. 5D) that is consistent with the observed effects on gene expression (G = Q allele = higher *PLAG1*/*CHCHD7* expression = more accessible; A = q allele = lower *PLAG1*/*CHCHD7* expression = less accessible). Moreover, the rs209821678 variant lies in a trough revealed in the ATAC-seq mode profile, suggesting that the corresponding segments mediated binding to a *trans*-acting factor. In this case, ATAC-seq data would therefore have been helpful in pinpointing the causative variants.

Discussion

We herein report the most complete catalog of open chromatin regions for cattle to date (Fig. 1; Supplemental File S1; e.g., Foissac et al. 2019; Halstead et al. 2020a,b; Kern et al. 2021; Ming et al. 2021). It comprises more than 976,000 ATAC-seq peaks detected in one or more of 63 tissue types representing pregastrulation embryos, endoderm, ectoderm, and mesoderm. To facilitate its use by the community, the data are made accessible via a custom track on the UCSC Genome Browser (via https://genome.ucsc.edu/s/Animal_Genomics_ULiege/ATAC_hub_V1 or https://www.gigauag.uliege.be/cms/c_4791343/en/gigauag-diagnostics-software-data). The vast majority of ATAC-seq peaks (about 840,000) show tissue-specific accessibility, dominated by one of 16 NMF components (weight of the largest NMF component >0.3). Of note, nearly 213,000 of these are specific for preimplantation embryonic stages. This clearly indicates that, as expected, chromatin accessibility is very dynamic, warranting the analyses of multiple tissues during fetal development in future studies. By studying across-tissue correlation between gene expression (using publicly available RNA-seq data) and accessibility of neighboring peaks, we show a clear signal of enhancer and/or promotor activity (excess of positive correlations with decreasing distance) but not of silencer activity (depletion of negative correlations with decreasing distance). This either indicates that silencers only account for a small minority of *cis*-acting regulatory elements or that silencers are not effectively identified using assays relying on chromatin openness (see also hereafter).

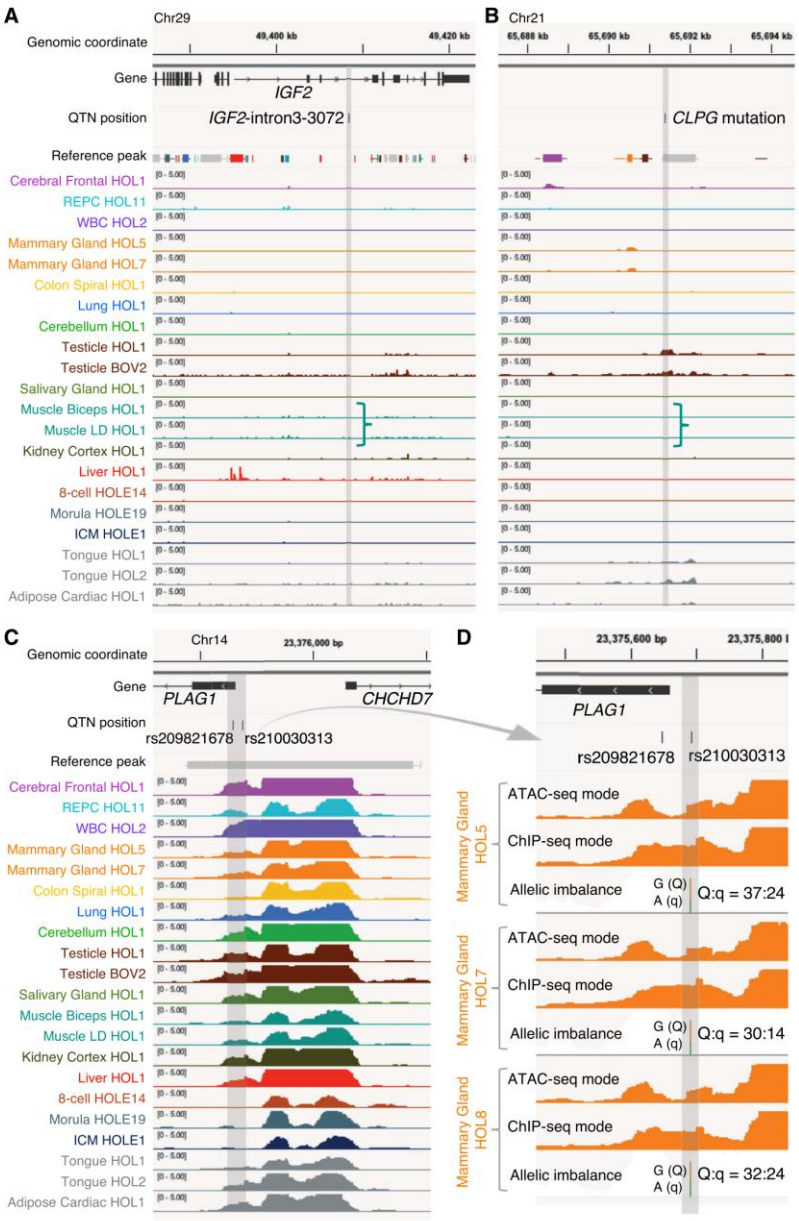


Figure 5. A retrospective evaluation of the utility of the ATAC-seq catalog for identifying regulatory variants. ATAC-seq peaks at three genomic loci encompassing regulatory QTNs previously identified in domestic animals. Chromosome coordinates, gene annotations, QTN positions, core and consensus reference peak regions (thick bars and horizontal lines, respectively; color-coded based on their highest NFM component), and peaks (ChIP-seq mode tag coverage unless otherwise mentioned) from at least one tissue sample representing each NMF component group with corresponding color code. Positions of the QTNs are highlighted as vertical gray bands. Track height measures the normalized tag coverage (1,000,000/[total tag count]). (*A*) The bovine orthologous region encompassing the *IGF2*-intron3-3072 QTN identified in pigs (A/G at susScr11: Chr 2: 1,483,817; bosTau9: Chr 29: 49,408,408) (Van Laere et al. 2003; Markljung et al. 2009) that maps to a 16-bp motif highly conserved among placental mammals disrupts interaction of the ZBED6 repressor, resulting in an approximately threefold up-regulation of *IGF2* in postnatal skeletal muscle affecting muscle growth, heart size, and fat deposition. None of ATAC-seq peaks overlapping the 16-bp motif were called across the 104 ATAC-seq data analyzed in this study. (*B*) The bovine orthologous region encompassing the callipyge (CLPG) muscular hypertrophy mutation identified in sheep (A/G at oviAri4: Chr 18: 64,294,536; bosTau9: Chr 21: 65,691,397) (Freking et al. 2002; Smit et al. 2003). The mutation is located in a 12-bp highly conserved motif among placental mammals and is considered to disrupt a muscle-specific long-range control element (a silencer) that causes ectopic expression of a 327-kb cluster of imprinted genes in postnatal skeletal muscle. ATAC-seq peaks overlapping the mutation site were called only in testis and tongue samples but not in skeletal muscle. (*C*) Bovine *PLAG1* promoter region encompassing two out of eight candidate QTNs influencing bovine stature identified by Karim et al. (2011) (rs209821678 [alternatively ss319607405]: (CCG)11/(CCG)9 at bosTau9: Chr 14: 23375648–23375650; rs210030313 [ss319607406]: G/A at bosTau9: Chr 14: 23375692). The two QTNs reside in a strong 1044-bp-long ubiquitous peak between the *PLAG1* and *CHCHD7* TSSs. Regions encompassing the other six credible variants do not map to any called peak in our ATAC-seq data and, hence, are not shown. (*D*) Enlargement of the two QTN loci for three animals that are Qq heterozygous at rs210030313. Peaks called with ATAC-seq and ChIP-seq modes, as well as allelic imbalance in mapped reads, are shown. The two QTNs reside in a footprint of the ATAC-seq mode peak, which is recovered by a ChIP-seq mode peak, indicating the presence of *trans*-acting factor(s) in the region hindering cleavage events by transposases. Allelic imbalance at rs210030313 (Q = G; q = A) indicates that the Q allele is more accessible compared with the q allele. Previous work showed that the two regulatory variants affect bidirectional promoter strength and that the Q allele, associated with bigger stature, showed approximately 1.5-fold higher promoter activity compared with the q allele in a luciferase assay. Figures were created using the Integrative Genomics Viewer (Robinson et al. 2011).

One of the main motivations to establish open chromatin catalogs in livestock is to identify regulatory variants that might underpin the heritability of agronomically important traits. Indeed, it is hoped that knowledge of these regulatory variants may increase the accuracy of genomic selection. We identified 1,390,391 variants mapping to open chromatin regions, of which 938,374 (67%) are common variants with MAF > 0.05 in Dutch HFs (Supplemental File S2). Instead of prioritizing variants mapping to ATAC-seq peaks indiscriminately for genomic selection, our catalog can be used to define sets of variants that are accessible in tissue types that are relevant for the trait under consideration. For instance, variants mapping to ATAC-seq peaks that are specifically accessible in, for instance, the mammary gland, hypothalamus, pituitary gland, and liver might be particularly relevant when targeting milk production traits.

We note that the proportion of SNVs (as opposed to indels) mapping to ATAC-seq peaks is significantly higher than the proportion of genome space occupied by ATAC-seq peaks (Fig. 3). We provide evidence that this is owing to an approximately 1.3-fold higher DNM rate in ATAC-seq peaks compared with the rest of the genome, corroborating recent findings in other eukaryotes (Kaiser et al. 2021; Luquette et al. 2022; Monroe et al. 2022). Shifts toward lower MAFs compared with variants in flanking regions support the operation of purifying selection on open chromatin regions and, hence, their functional importance (Fig. 3).

To further examine the regulatory function of open chromatin regions, we identified 7817 and 6172 sets of credible variants assumed to include causative variants driving the same number of *cis* eQTLs detected in blood and liver, respectively (Supplemental Table S12). As anticipated, variants in these credible sets tend to map to open chromatin regions more often than expected by chance alone (as evaluated by permutation following the method of Trynka et al. 2015). Furthermore, the enrichment was not random with respect to the NMF component (Fig 4; Supplemental Table S13). Credible sets corresponding to blood eQTLs tended to overlap ATAC-seq peaks assigned to the immune and ubiquitous NMF, whereas liver eQTLs tended to overlap ATAC-seq peaks assigned to the liver and ubiquitous NMF. The overlap with the tissue-specific NMF (immune and liver) was primarily owing to distant regulatory elements, whereas the overlap with the ubiquitous NMF was equally owing to distant and proximal regulatory elements (Fig. 4).

These results tell us that variants that map to ATAC-seq peaks are more likely to be regulatory variants than variants that map outside of ATAC-seq peaks. However, they do not really tell us how sensitive and precise ATAC-seq assays are to identify regulatory variants. We summarized this interrogation with two specific questions: (1) what fraction of regulatory variants map to ATAC-seq peaks (sensitivity), and (2) what fraction of variants in ATAC-seq peaks are regulatory (precision). We developed a maximum likelihood framework using eQTL information to estimate both parameters. Sensitivity was estimated at one in three, and precision at one in 25. Thus, as many as two out of three regulatory variants may lie outside of ATAC-seq peaks inventoried in our catalog (Supplemental File S2). A first possible explanation of this observation is that our catalog still misses a substantial proportion of bovine gene switches. This could be because, in particular, we did not explore sufficient developmental stages, or ATAC-seq peaks do not capture all gene switches (e.g., silencers). A second possible explanation is that variants lying outside of ATAC-seq peaks may nevertheless affect the functionality of (nearby) gene switch components that are identified by ATAC-seq peaks (e.g., by affecting the formation of secondary structures involving the switch). Finally, some variants are known to affect transcript levels not by perturbing gene switches but by affecting transcript stability, including stop gains and splice variants. For example, the K232A mutation in *DGAT1* affects transcript abundance by affecting splicing (Fink et al. 2020). The one in 25 precision indicates that the majority of variants falling in ATAC-seq peaks are probably neutral. It is possible that some eQTLs are driven by more than one causative variant, which would slightly increase precision. Contrary to coding variants, which can be identified quite accurately based on our understanding of the genetic code and splicing mechanisms, predicting the effect of SNVs on the functionality of *cis*-acting regulatory elements is still in its infancy.

Taken together, our results indicate that the knowledge of open chromatin regions in the bovine genome is a first step toward the identification of regulatory variants, yet this knowledge will likely have to be complemented with additional information to more effectively pinpoint the causative regulatory variants and thereby have a major impact on the accuracy of genomic selection.

Methods

Ethical approvals, sample collection, and processing

All relevant procedures using animals were approved by the animal care and use committee (ACUC) of the University of Liège (approval no. 17-1948 and 17-1949) or the Ruakura ethics committee, Hamilton, New Zealand (approval no. AEC 12845) and performed in accordance with the relevant guidelines and regulations of the committees. Blood samples were collected from the tail or jugular vein of animals using K2-EDTA blood collection tubes. White blood cells (WBCs) were enriched by lysing erythrocytes using eBioscience 10X RBC lysis buffer (Thermo Fisher Scientific). Peripheral blood mononuclear cells (PBMCs) were prepared by density gradient centrifugation using either Ficoll-Paque plus (Cytiva) or Lymphoprep (Stemcell Technologies) with SepMate-50 tubes (Stemcell Technologies). For tissue collections, animals were humanely euthanized, and tissues were collected in ice-cold Belzer UW cold storage solution (Bridge to Life) until processing as described below or were otherwise snap-frozen in liquid nitrogen. Details of bovine samples can be found in Supplemental Tables S1 and S2. To optimize protocols, one C57BL/6J × A/J F1 mouse (male, 1 yr old) was euthanized by cervical dislocation. Four tissues (liver, spleen, kidney, muscle) were collected in the UW cold storage solution and processed as described below. We processed/stored samples in three ways: fresh, slow-frozen, and snap-frozen conditions.

Fresh samples

Tissues collected in the UW cold storage solution were directly used for constructing ATAC-seq libraries on the day of sampling (17 biosamples with "fresh" in their names) (Supplemental Table S3).

Cryopreserved samples

Tissues were cut into ~27-mm^3 cubes and transferred to cryotubes filled with 1 mL of STEM-CELLBANKER DMSO-free cell freezing media (Amsbio). The tubes were kept on ice for ~10 min and transferred to a cryo-box in dry ice while other samples were processed. The samples were then stored in a −80°C freezer until use (55 biosamples with "slow" in names) (Supplemental Table S3).

Snap-frozen samples

Tissues were cut into ~27-mm^3 cubes and transferred to cryotubes. The samples were snap-frozen in liquid nitrogen and stored at −80°C until use (34 biosamples with "snap" in names) (Supplemental Table S3).

ATAC-seq library construction

ATAC-seq libraries were constructed following the Omni ATAC-seq protocol (Corces et al. 2017) with some modifications.

Tissue homogenization

A cryopreserved tissue in a vial was quickly thawed in a water bath and transferred to an excess amount of ice-cold Dulbecco's phosphate buffered saline (DPBS). The cryopreserved or fresh tissue samples were dissociated into a single-cell suspension using a gentleMACS dissociator (Miltenyi Biotec) by running one or two cycles of program B1 with 3 mL of ice-cold Omni 1 × homogenization buffer in a gentleMACS C-tube. Snap-frozen samples were pulverized using a mortar and pestle chilled in liquid nitrogen. The cell suspension or pulverized tissue was transferred to a Dounce tissue grinder (Merck D9063) on ice with 3 mL of ice-cold Omni 1 × homogenization buffer. Samples were homogenized with an A-pestle until resistance went away and further with a B-pestle (three to 10 strokes each) so as to disrupt cellular plasma membranes. Cell debris were removed by passing the sample through stackable cell strainers (100-, 70-, and 30-µm MACS SmartStrainers, Miltenyi Biotec). The flow-through was further clarified by a brief centrifugation at 100*g* for 1 min at 4°C. The supernatant was mixed with an equal volume of ice-cold Omni 50% iodixanol solution (final, 25% iodixanol).

Purification of nuclei

Two layers of iodixanol cushions were prepared in a 2-mL LoBind tube (Eppendorf) by placing 600 µL of ice-cold 40% iodixanol solution on the bottom (marking the surface of the bottom layer facilitated sample collection later) and overlaying 600 µL of ice-cold 29% iodixanol solution using a wide-bore tip. On the top, 800 µL of the cell suspension (containing 25% iodixanol) was placed. Density gradient centrifugation was performed using a table-top centrifuge (Eppendorf 5430R) with a swing rotor at 6000*g* for 30 min at 4°C and a soft brake setting. Top layers were carefully removed down to ~2 mm above the bottom layer. The nuclear fraction, between the bottom and middle layers, was collected (~400 µL) to a new LoBind tube on ice. The number of nuclei was counted by mixing 20 µL of the sample with 20 µL of 100 × diluted Hoechst 33342 (Thermo Fisher Scientific) using a hemocytometer under a fluorescence microscope.

Tagmentation

Approximately 50,000 nuclei were transferred to two 1.5-mL LoBind tubes filled with 1 mL of ice-cold Omni-ATAC-RSB + 0.1% Tween-20 buffer and centrifuged at 500*g* for 10 min at 4°C. After carefully removing the supernatant, nuclei were resuspended in 50 µL of Omni-ATAC reaction mix containing Tn5 transposase TDE1 enzyme (Illumina). As the effectiveness of transposase varied slightly among samples, we used two different amounts of the enzyme per sample (Supplemental Table S3). After mixing the sample by pipetting six times using a P200 fine tip, tagmentation reaction was performed using an Eppendorf ThermoMixer at 500 rpm for 30 min at 37°C. The reaction was stopped by adding 300 µL of PB buffer in a MinElute PCR purification kit (Qiagen) and 10 µL of 3 M sodium acetate (pH 5.2). The sample was mixed, kept at room temperature for 10 min, and stored at −20°C until DNA purification. Libraries for two blood samples (WBC, PBMC) were generated with an alternative ATAC-seq protocol (Buenrostro et al. 2013) during the pilot experiment phase. A genomic DNA (gDNA) control library was also prepared using 50 ng of purified gDNA from one animal (HOL1_m) by following the Nextera DNA sample preparation guide (Illumina). The tagmented DNA was purified and eluted in 21 µL elution buffer using the MinElute PCR purification kit.

Library preparation

The purified DNA was amplified using NEBNext high-fidelity 2X PCR master mix with the Ad1 and Ad2 primers (Buenrostro et al. 2013) for 13 (for ATAC-seq library) or five PCR-cycles (gDNA library), respectively. The amplified libraries were purified and eluted in 50 µL elution buffer using the MinElute PCR purification kit. Library size distribution was monitored using 10 µL of the library by QIAxcel capillary electrophoresis (Qiagen). Large DNA fragments were eliminated using AMPure XP magnetic beads (Beckman Coulter) by a right-side size selection using 0.55 × followed by 1.5 × volume ratio of beads to sample. Library concentration was estimated using the KAPA library quantification kit (Kapa Biosystems). The libraries were sequenced with 2 × 38-bp paired-end reads using a NextSeq 500 sequencer, or 2 × 51-bp paired-end reads on a NovaSeq 6000 (Illumina) instrument. In total, 185 ATAC-seq libraries were sequenced, yielding 31.8 million paired-end fragments on average (range: 2.5–117.2 million fragments) (Supplemental Table S3). ATAC-seq FASTQ files were obtained from the EMBL-EBI ArrayExpress (https://www.ebi.ac.uk/biostudies/arrayexpress) under accession number E-MTAB-9872 (Lee et al. 2021) or generated in this study and submitted under accession numbers E-MTAB-11825 and E-MTAB-11826 (see Data access). In addition, we downloaded publicly available ATAC-seq data from the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA531214, PRJNA665194, PRJNA601200, PRJNA595394, and PRJNA622966 (Supplemental Table S1; Fang et al. 2019; Halstead et al. 2020a,b; Johnston et al. 2021) and analyzed these in a similar way.

ATAC-seq peak calling

ATAC-seq peaks were called following the recommendations of the ENCODE ATAC-seq pipeline ("ATAC-seq Data Standards and Processing Pipeline") (https://www.encodeproject.org/atac-seq/).

Trimming

Sequences with low sequence quality, residuals of library adaptors, and bases >38 bp (to uniform read length across data) were trimmed using Trimmomatic (ILLUMINACLIP:NexteraPE-PE.fa:2:30:5:1:true SLIDINGWINDOW:4:15 MINLEN:20 CROP:38) (Bolger et al. 2014). Proportions of reads remaining after trimming averaged 98.7% (range: 96.8–99.5%) (Supplemental Table S3).

Mapping

The trimmed reads were aligned to the bovine genome assembly ARS-UCD1.2 using Bowtie 2 (--local –mm). Overall mapping rate averaged 95.5% (range: 36.4%–99.4%).

Filtration

Reads mapping to the mitochondrial chromosome were filtered out using SAMtools (samtools idxstats file.bam | cut -f 1 | grep -v chrM | xargs samtools view -b file.bam) (Danecek et al. 2021). The proportion of mitochondrial reads averaged 16.1% (range: 0.8%–61.9%).

PCR/optical duplicates were removed using Picard toolkit (java -jar picard.jar MarkDuplicates REMOVE_DUPLICATES = true OPTICAL_DUPLICATE_PIXEL_DISTANCE = 100 [2500 for NovaSeq data] VALIDATION_STRINGENCY = LENIENT) (http://broadinstitute.github.io/picard/). Duplicate read rate averaged 12.7% (range: 4.9%–23.1%). Properly aligned reads with high sequencing quality were selected with SAMtools (for paired-end reads, samtools view -f 3 -F 1284 -q 30; for single-end reads after trimming, samtools view -f 9 -F 260 -q 30), resulting in an average of 35.3 million informative reads per library (range: 3.4–141.5 million).

Fractionation

Reads were partitioned into two bins based on their mapped fragment lengths using BamTools filter function (Barnett et al. 2011): (1) short reads generated from putative nucleosome-free regions of DNA (<146 bp) and (2) longer reads likely from nucleosome-associated DNA (≥146 bp).

Peak calling

ATAC-seq peaks were called using MACS2 in two ways (Supplemental Fig. S3): First, in ATAC-seq mode: the genomic locus cleaved by the transposase (a tag) was defined as a 38-bp region centered either 4 bp (for a plus strand read) or 5 bp (for a minus strand read) downstream from the read's 5′-end (Adey et al. 2010; Buenrostro et al. 2013). Peaks (open chromatin regions) were identified by comparing the tag distribution of a sample to one from a purified gDNA control (macs2 callpeak --format BED --control --nomodel --shift -19 --extsize 38 --qvalue 0.05 --gsize hs --keep-dup all --max-gap 38 --SPMR --bdg). Second, in ChIP-seq mode, to recover regions protected from transposase cleavage events owing to binding of *trans*-regulatory factor(s) like a TF (so-called footprints in the ATAC-seq mode analysis), peaks were called using general settings used for ChIP-seq analysis (MACS2 piles up entire sequencing fragments instead of focusing on transposase cleavage sites close to 5′-ends of reads). To avoid covering nucleosome positions, only the nucleosome-free fraction of sequence fragments (mapped fragment size < 146 bp) was used (macs2 callpeak --format BAMPE --control --qvalue 0.05 --gsize hs --keep-dup all --max-gap 38) for ChIP-seq mode.

Quality control

In-house ATAC-seq data (peak-called with the ATAC-seq mode) were evaluated using commonly used ATAC-seq quality-control measurements (Supplemental Table S3): the fraction of reads in called peak regions (FRiP; average: 0.212; range: 0.004–0.591) and TSS enrichment (average: 16.1; range: 2.14–46.2). Irreproducible discovery rates (IDRs) that measure reproducibility in score ranking between peaks, as well as rescue ratios that measure consistency between replicates (average: 1.23; range: 1.02–1.96), were calculated using samples with technical replicates.

Low-quality libraries with the number of filtered reads of fewer than 10 million per sample, FRiP less than 0.07, TSS enrichment less than 7.0, and a self-consistency ratio more than two, as well as some technical duplicates with less quality, were excluded from further analyses (36 libraries out of 185).

Final peak calling per biosample

Reads of libraries from the same biosample that passed the quality control were merged, and peaks were called afresh as described above in ATAC-seq and ChIP-seq modes. We also integrated 15 high-quality public data sets in our analysis (Fang et al. 2019; Halstead et al. 2020a,b; Johnston et al. 2021). *P*-value thresholds for final peak selection for all samples (with and without technical replicates) were determined as the median of the lowest log(1/*P*)-values of peaks with IDR ≤ 0.1 across samples with technical replicates ($-\log_{10}$(*P*-value) = 8.01 and 9.28 for ATAC-seq and ChIP-seq modes, respectively).

Reproducibility

Reproducibility of peak calling was evaluated by measuring Pearson's correlation of genome-wide read coverage in 500-bp windows between technical (range: 0.89–0.99) and biological replicates (0.85–0.97) (Supplemental Figs. S1, S2) using deepTools' bamCoverage (--outFileFormat bigwig --effectiveGenomeSize 2701495761 --normalizeUsing RPKM --ignoreForNormalization chrX chrY), multiBigwigSummary (bins --binSize 500), and plot Correlation (--corMethod pearson --whatToPlot heatmap –remove Outliers --colorMap viridis --plotNumbers) (Ramírez et al. 2016).

Defining and merging consensus and core peak components across samples and calling modes

Core and consensus peak components were defined following the method of Meuleman et al. (2020) as follows. Individual peak (IP) summit positions were collated across samples separately for ATAC-seq and ChIP-seq modes. Summits were clustered such that the distance between clusters was ≥20 bp. The space covered by each cluster was defined as the core component of a newly defined "collective peak" (CP). The corresponding IP were piled up and the limits of the consensus CP defined as the full-width at half maximum. If by doing so some consensus CPs overlapped, they were merged by repeating the process using all concerned IPs. This yielded two genome-wide sets of core and consensus CPs for ATAC-seq mode and ChIP-seq mode, respectively. If overlapping, the corresponding consensus and core CPs were merged to, in the end, yield one unique set of core and consensus "CPs" or reference peaks used for all further analyses (Supplemental File S1).

Nonnegative matrix factorization

NMF was conducted following the method of Meuleman et al. (2020) using scripts downloaded from GitHub (https://github.com/Altius/Vocabulary). Briefly, we set up an m (number of samples) × n (number of peaks) matrix ($\mathbf{V}$) summarizing the accessibility of each peak in each sample in binary mode (0 or 1, based on presence absence of an "IP" in the corresponding sample). $\mathbf{V}$ was decomposed in a $m \times k$ $\mathbf{W}$ and $k \times n$ $\mathbf{H}$ matrix such that $\mathbf{V} \approx \mathbf{W} \times \mathbf{H}$, where k is the number of hidden components. The value for k was set at 16 following the method of Meuleman et al. (2020), as a trade-off between maximizing the recapitulation of $\mathbf{V}$ and retaining biological interpretability (k at elbow point of the derivative of F1 score over k). Following this procedure, each sample and each peak were assigned a weight for each one of the k components. Samples and peaks were in general assigned to their dominant component (with largest score), and components were assigned to biological systems on the basis of their composite samples (Fig. 1F; Supplemental Table S5; Supplemental File S1). The peak information is also made accessible via a custom track on the UCSC Genome Browser (https://genome.ucsc.edu/s/Animal_Genomics_ULiege/ATAC_hub_V1 or https://www.gigauag.uliege.be/cms/c_4791343/en/gigauag-diagnostics-softwar e-data).

Public RNA-seq data

RNA-seq data originated from bovine tissues similar to the ones generated in this study were downloaded from the NCBI Sequence Read

Archive (https://www.ncbi.nlm.nih.gov/sra; for accession numbers, see Supplemental Table S9; Graf et al. 2014; Cai et al. 2018; Dado-Senn et al. 2018; Khansefid et al. 2018; Fang et al. 2019, 2020). Low-quality bases, residuals of library adaptors, and short reads <35 bp were removed using Trimmomatic (ILLUMINACLIP:adaptor.fa:2:30:10:4:true SLIDINGWINDOW:4:15 LEADING:10 TRAILING:10 MINLEN:36). The reads were mapped to the bovine reference genome ARS-UCD1.2 using HISAT2 (version 2.1.0) using genome indexes that were built along with coordinates of 2.7 million DNA variants (2,389,896 SNVs and 176,799 indels) and reference transcripts (bosTau9.ncbiRefSeq.gtf; hisat2 --dta --no-softclip -x index -S out.sam). Reads mapped to ribosomal RNA, duplicated, or improperly mapped were filtered out using BEDTools (intersectBed -abam -b rRNA.bed -v), Picard toolkit (MarkDuplicates REMOVE_DUPLICATES = true VALIDATION_STRINGENCY = LENIENT), and SAMtools (view -f 3 -F 1284 -q 30). Expression levels of reference transcripts on autosomes and sex chromosomes in the gene reference file (bosTau9.ensGene.gtf, v101) were estimated using StringTie (stringtie -G -e). A raw read count matrix per gene (27,233 genes) was prepared using a prepDE.py script in StringTie. Using R package DESeq2, the count data were normalized by their library sizes after selecting genes for which counts were more than 10 (for the TF binding motif enrichment analysis) or 30 (for the peak-expression correlation analysis) in at least one sample (Data < –Data [(rowSums (counts (Data)) > threshold) ≥ 1]) and transformed using regularized-logarithm transformation (rlog (Data, blind = TRUE)).

TF binding motif enrichment in peaks

Known and de novo DNA motifs enriched in core peaks assigned to tissue specific components (one weight ≥ 0.9) or the ubiquitous component (all weights ≤ 30%) were identified using HOMER (findMotifsGenome.pl peak.bed bosTau9_genome_directory-size given) (Fig. 2A; Supplemental Tables S7, S8). For each of the 16 components, we then checked, for the 10 most enriched binding motifs, whether the cognate TF was also more strongly expressed in the tissue samples assigned to that component. This was accomplished by standardizing (*Z*-score) the expression level of the corresponding TF gene across 114 of the above-mentioned publicly available RNA-seq libraries that could be assigned to one of our 16 components (Supplemental Table S7), and verifying whether the *Z*-scores were higher in the tissue type assigned to the cognate component compared with the other samples. The statistical significance of the difference in *Z*-score was estimated using a permutation test and ensuing *P*-values converted to FDR values (π_0 set at 1) using the qvalue R package (http://github.com/jdstorey/qvalue) to correct for multiple testing. FDRs ≤ 0.05 were deemed significant.

Correlation between chromatin openness and gene expression

Chromatin openness

Chromatin openness of a peak in a given sample was measured as the fold enrichment (in normalized read depth) over gDNA background at the nucleotide position in the peak with the highest such value. This was computed with the MACS2 bdgcmp function (-m FE -t sample_pileup.bdg -c control_lambda.bdg) and using the bedGraph files from MACS2 ATAC-seq mode peak calling. The highest fold enrichment value per peak was extracted using BEDTools (map -nonamecheck -c 4 -o max -a consensus_peak.bed -b out.bdg). We kept peaks for which fold enrichments were more than five at least in one sample (975,488 peaks).

Gene expression

Fifty-six bovine public RNA-seq data matched to 91 of our 104 ATAC-seq data sets were selected from the data sets mentioned above (Supplemental Table S9). RNA-seq data were processed as described above.

Correlation

Pearson's correlations between openness (fold enrichment) of peaks located within 1 Mb from TSS of a given gene and gene expression level across the 91 data sets were calculated using R stats (R Core Team 2023).

Generating a catalog of common variants mapping to *cis*-acting regulatory elements

Genome-wide variant catalog

We used a catalog of 11,030,905 SNVs and 1,705,738 short (≤265 bp) indels called from whole-genome sequences of 264 HF cattle (obtained from BioProject accession number PRJEB53518; Oget-Ebrad et al. 2022).

Proportion of variants mapping to ATAC-seq peaks by genome-compartment

The genome was subdivided in five mutually exclusive compartments (TSS, TTS, exonic, intronic, intergenic) using a gene reference file (bosTau9.ensGene.gtf, v101). Each compartment was further subdivided in (1) a part overlapping any peak in our catalog of 948,566 autosomal consensus peaks and (2) the remaining part. We then checked whether there was a significant difference in the proportion of variants mapping to the peak part versus the proportion of space occupied by the peak part using a chi-square goodness-of-fit test.

Density of singletons within and outside of ATAC-seq peaks

The change of singleton density as a function of the distance from the nearest peak was determined sequentially as follows. We first identified the size of the genome (in base pairs) that was within 100 bp from the nearest ATAC-seq peak (200-bp windows centered on the peaks; g_1), as well as the number of singletons that mapped within this space (s_1), and computed the corresponding ratio ($r_1 = s_1/g_1$). We then identified the size of the genome that was between 300 and 100 bp from the nearest ATAC-seq peaks (200-bp windows on the left and right of window 1, excluding what was assigned to fraction 1; g_{2L} and g_{2R}), as well as the number of singletons that mapped within these spaces (s_{2L} and s_{2R}), and computed the corresponding ratios ($r_{2L} = s_{2L}/g_{2L}$ and $r_{2R} = s_{2R}/g_{2R}$). We pursued this process for windows that were more and more distant from the nearest ATAC-seq peaks. The "confidence interval" around the estimates was defined as the computed ratio ±2 SD ($2 \times SD_i$), where SD_i was computed assuming a binomial distribution as $SD_i = \sqrt{g_i \times r_i \times (1 - r_i)}$.

Site frequency spectrum

Variants mapping to peaks were sorted according to the five above-mentioned compartments (TSS, TTS, exonic, intronic, intergenic). Their SFS (0.01 bins) was compared (histogram) between compartments and with that of all other variants in the genome. To check for a shift toward lower allelic frequencies by compartment, we compared the distribution of allelic frequencies between variants that mapped to peaks ("peak part" above) versus variants that did not map to peaks (but belonged to the same compartment) using a Wilcoxon rank-sum test.

Identifying bovine *cis* eQTLs in liver and blood

RNA-seq and data preprocessing

We reanalyzed RNA-seq data of 176 liver and 227 whole-blood biopsies collected from 240 Holstein females at ~14 d postpartum in the GplusE project (obtained from ArrayExpress; accession numbers E-MTAB-9347 and E-MTAB-9431 for blood; E-MTAB-9348 and E-MTAB-9871 for liver) (Lee et al. 2021; Wathes et al. 2021a,b). The libraries were constructed with an Illumina TruSeq stranded total RNA library prep Ribo-Zero gold kit and sequenced with 75-base single-end reads. First, low-quality bases, residuals of library adaptors, and short reads (<35 bp) were removed using Trimmomatic (java -jar trimmomatic-0.36.jar SE input.fastq.gz output_trimmed.fastq.gz ILLUMINACLIP:TruSeq3-SE.fa:2:30:10 SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3 MINLEN:36 2>>log.txt). The reads were mapped to the bovine reference genome ARS-UCD1.2 using HISAT2 (hisat2 --dta --no-softclip -x index -U trimmed.fastq.gz -S output.sam --rna-strandness R 2>>log.txt). Reads mapped on ribosomal RNA were filtered out using SAMtools (samtools sort input.sam -o sorted.bam) and BEDTools (intersectBed -abam sorted.bam -b rRNA.bed -v). BAM files from the same biosample were merged, and properly mapped reads were kept with SAMtools (samtools merge merged.bam sample_ID*.bam; samtools view -F 2308 -q 30 -o clean.bam -b merged.bam; samtools sort -o clean.sorted.bam clean.bam; samtools index clean.sorted.bam). Expression levels of reference transcripts (bosTau9.ensGene.gtf, v101) on autosomes and sex chromosomes were estimated using StringTie (stringtie clean.sorted.bam --rf -G bosTau9.ensGene.nochrUn.gtf -e -o transcripts.gtf) (Pertea et al. 2015). A raw read counts matrix by gene (27,233 genes) was prepared using a prepDE.py script in StringTie. Gene-specific reads counts were scaled with a "size factor" using DESeq2 after eliminating mitochondrial gene counts. Gene with summation of TPM lower than one and with fewer than eight individuals with counts greater than zero were filtered out. Afterward, counts were normalized by inverse normal transformation by gene and across individuals.

SNV genotyping

All animals were genotyped with a high-density (about 778,000) SNV array (Illumina BovineHD Genotyping BeadChip), and imputed to whole genome using SHAPEIT4 (for phasing) (Delaneau et al. 2019) and Minimac4 using the previously mentioned whole-genome sequences of 264 HF animals as reference. Variants with MAF ≤ 0.02, probability of the data assuming Hardy–Weinberg equilibrium ≤0.001, and imputation accuracy (r^2) ≤0.9 were filtered out, leaving a total of 10,257,878 usable markers (the genotypes are available from the Zenodo open data repository at https://doi.org/10.5281/zenodo.8339268).

eQTL analyses

We used QTLtools to ensure RNA–DNA sample matching based on genotype concordance (Supplemental Table S11). Expression values were first corrected for hidden confounders and "country of origin" using probabilistic estimation of expression residuals (PEER). The resulting residuals were then further corrected for stratification and/or polygenic effects on gene expression using GenABEL. The ensuing "double-corrected" residuals were then used for *cis* eQTL analyses using QTLtools. For each gene, we tested all variants within 1 Mb from the TSS. Ensuing *P*-values were corrected for multiple testing (in the window) by permutation. For each gene, we kept the best *P*-value (= "lead variant"), and these "best *P*-values" were converted to FDR and *Q*-values (hence corrected for multiple testing) following the methods of Benjamini and Hochberg (1995) and Storey and Tibshirani (2003), respectively. eQTL with FDR ≤ 0.05 were deemed experiment-wide significant. π_1 (the proportion of alternative hypotheses among all tested hypotheses) was estimated according to the method of Storey and Tibshirani (2003).

Enrichment of eQTL driving variants in ATAC-seq peaks

We first identified, for each significant *cis* eQTL, a credible set of variants defined as all the variants within 1 Mb from the lead variant and in LD with it with a threshold value $r^2_{LD} \geq 0.9$. We then used the method proposed by Trynka et al. (2015) to measure the putative enrichment of credible variants in ATAC-seq peaks. The analysis was performed by NMF component. Briefly, we defined, for each *cis* eQTL, a region/window spanned by the credible set plus buffer segments on either side corresponding to twice the median peak size (= 436 bp). We first counted, using the real eQTL results, for how many eQTLs at least one credible variant mapped within an ATAC-seq peak (assigned to the NMF component under consideration). We then randomly shifted variant and peak coordinates with respect to each other within each *cis* eQTL window and counted for how many eQTLs at least one credible variant mapped within an ATAC-seq peak. This "permutation" process was repeated 10,000 times, and the significance of the overlap between credible variants and peaks observed for the real data was evaluated from the number of occurrences of an equal or higher overlap with the permuted data. To evaluate the contribution of proximal rather than distal ATAC-seq peaks to the signal, permutations were also conducted separately by peak type (proximal vs. distal).

Estimating the proportion of regulatory variants mapping in ATAC-seq peaks and the proportion of variants mapping in ATAC-seq peaks that are regulatory

We assumed that every *cis* eQTL i out of T is driven by one regulatory variant that is part of a credible set comprising n_{iA} variants in the ATAC-seq peaks and n_{iN} variants outside of the ATAC-seq peaks. We further assumed that a fraction f_A of *cis* eQTLs is driven by a regulatory variant mapping to an ATAC-seq peak (Supplemental Codes S1, S2), as well as a fraction $f_N = 1 - f_A$ by a regulatory variant mapping outside ATAC-seq peaks, and that ATAC-seq peaks occupy a proportion p_A of the genome. The likelihood of the data for eQTL i can hence be expressed as

$$L_i = f_A\left[\binom{n_{iA}+n_{iN}-1}{n_{iA}-1} \times {p_A}^{n_{iA}-1} \times (1-p_A)^{n_{iN}}\right] + f_N\left[\binom{n_{iA}+n_{iN}-1}{n_{iN}-1} \times {p_A}^{n_{iA}} \times (1-p_A)^{n_{iN}-1}\right].$$

This equation assumes that the $(n_{iA}+n_{iN}-1)$ "passenger" variants in the credible set are distributed between ATAC-seq peaks and the rest of the genome according to the proportion of the genome occupied by these two components and following a binomial distribution.

We used the Newton–Raphson method (R nlm function) (R Core Team 2023) to determine the value of f_A that maximizes the likelihood of the data for all T eQTL:

$$L_T = \prod_{i=1}^{T} L_i.$$

f_A corresponds to the above-mentioned sensitivity (*s.t.* $0 \leq f_A \leq 1$), whereas the precision was estimated as

$$\frac{f_A \times T}{\sum_{i=1}^{T} n_{iA}}.$$

Data access

ATAC-seq data generated in this study have been submitted to the EMBL-EBI ArrayExpress (https://www.ebi.ac.uk/biostudies/arrayexpress) under accession numbers E-MTAB-11825 and E-MTAB-11826. Imputed genotypes of animals used for eQTL analyses are available from the Zenodo open data repository at https://doi.org/10.5281/zenodo.8339268. Other data sets used in this study, published previously, are described in the Methods. Key analysis pipelines are available at GitHub (https://github.com/can11sichuan/Bov-ATAC) and Supplemental Code S1 and S2. The UCSC Genome Browser track hub to visualize all 104 individual and reference ATAC-seq peaks is accessible from https://genome.ucsc.edu/s/Animal_Genomics_ULiege/ATAC_hub_V1 or https://www.gigauag.uliege.be/cms/c_4791343/en/gigauag-diagnostics-software-data. Bovine ATAC-seq peaks and putative regulatory variants identified in this study are found in Supplemental Files S1 and S2, respectively.

GplusE Consortium[9]

Mark Crowe, Niamh McLoughlin, Alan Fahey, Elizabeth Matthews, Andreia Santoro, Colin Byrne, Pauline Rudd, Roisin O'Flaherty, Sinead Hallinan, Claire Wathes, Zhangrui Cheng, Ali Fouladi, Geoff Pollott, Dirk Werling, Beatriz Sanz Bernardo, Mazdak Salavati, Laura Buggiotti, Alistair Wylie, Matt Bell, Conrad Ferris, Mieke Vaneetvelde, Kristof Hermans, Geert Opsomer, Sander Moerman, Jenne De Koster, Hannes Bogaert, Jan Vandepitte, Leila Vandevelde, Bonny Vanranst, Johanna Hoglund, Susanne Dahl, Klaus Ingvartsen, Martin Sørensen, Leslie Foldager, Soren Ostergaard, Janne Rothmann, Mogens Krogh, Else Meyer, Charlotte Gaillard, Jehan Ettema, Tine Rousing, Federica Signorelli, Francesco Napolitano, Bianca Moioli, Alessandra Crisa, Luca Buttazzoni, Jennifer McClure, Daragh Matthews, Francis Kearney, Andrew Cromie, Matt McClure, Shujun Zhang, Xing Chen, Huanchun Chen, Junlong Zhao, Liguo Yang, Guohua Hua, Chen Tan, Guiqiang Wang, Michel Bonneau, Andrea Pompozzi, Armin Pearn, Arnold Evertson, Linda Kosten, Anders Fogh, Thomas Andersen, Matthew Lucy, Chris Elsik, Gavin Conant, Jerry Taylor, Nicolas Gengler, Michel Georges, Frederic Colinet, Marilou Ramos Pamplona, Hedi Hammami, Catherine Bastin, Haruko Takeda, Aurelie Laine, Lijing Tang, Martin Schulze, Cinzia Marchitelli, and Sergio Palma-Vera

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Calixte Bayrou, Ken Kusakabe, Ruth Appeltant, Anne-Sophie Van Laere, and all members of Michel Georges' laboratory for their help for sample collections, technical support, and fruitful discussion. We also thank the support provided by the GIGA Genomics and Bioinformatics core facilities. This work was funded by the Damona European Research Council advanced grant from the EU (AdG-GA323030), the GplusE FP7 grant from the EU (no. 613689), the CAUSEL grant from the Walloon Region (no. 1710030), and financial support from Inoveo. C.C. and T.D. are senior research associate and research director from the Fonds de la Recherche Scientifique. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif, funded by the Fonds de la Recherche Scientifique de Belgique (no. 2.5020.11) and by the Walloon Region.

Author contributions: H.T., M.G., T.D., C.C., D.C.W., and M.A.C. conceived and supervised the project. G.C., A.S., G.C.M.M., C.C., Z.C., M.S., and M.L. contributed to sample and data collections. H.T. and L.T. performed wet-laboratory experiments. V.A.P., C.O.-E., G.C.M.M., J.L.G., T.L., W.C., and T.D. assisted data analysis. C.Y., H.T., and M.G. performed data analysis. M.G., H.T., and C.Y. wrote the manuscript.

References

Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, et al. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* **11:** R119. doi:10.1186/gb-2010-11-12-r119

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11:** R106. doi:10.1186/gb-2010-11-10-r106

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23:** 1294–1296. doi:10.1093/bioinformatics/btm108

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27:** 1691–1692. doi:10.1093/bioinformatics/btr174

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc* **57:** 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120. doi:10.1093/bioinformatics/btu170

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10:** 1213–1218. doi:10.1038/nmeth.2688

Cai W, Li C, Liu S, Zhou C, Yin H, Song J, Zhang Q, Zhang S. 2018. Genome wide identification of novel long non-coding RNAs and their potential associations with milk proteins in Chinese Holstein cows. *Front Genet* **9:** 281. doi:10.3389/fgene.2018.00281

Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15:** 901–913. doi:10.1101/gr.3577405

Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B, et al. 2017. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14:** 959–962. doi:10.1038/nmeth.4396

Dado-Senn B, Skibiel AL, Fabris TF, Zhang Y, Dahl GE, Peñagaricano F, Laporta J. 2018. RNA-seq reveals novel genes and pathways involved in bovine mammary involution during the dry period and under environmental heat stress. *Sci Rep* **8:** 11096. doi:10.1038/s41598-018-29420-8

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10:** giab008. doi:10.1093/gigascience/giab008

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. 2016. Next-generation genotype imputation service and methods. *Nat Genet* **48:** 1284–1287. doi:10.1038/ng.3656

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6:** e1001025. doi:10.1371/journal.pcbi.1001025

Delaneau O, Ongen H, Brown AA, Fort A, Panousis NI, Dermitzakis ET. 2017. A complete tool set for molecular QTL discovery and analysis. *Nat Commun* **8:** 15452. doi:10.1038/ncomms15452

Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. 2019. Accurate, scalable and integrative haplotype estimation. *Nat Commun* **10:** 5436. doi:10.1038/s41467-019-13225-y

The ENCODE Project Consortium, Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Ai R, Aken B, Akiyama JA, Jammal OA, et al. 2020.

[9]School of Veterinary Medicine, University College Dublin, Dublin 4, Ireland

Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583:** 699–710. doi:10.1038/s41586-020-2493-4
Fang L, Liu S, Liu M, Kang X, Lin S, Li B, Connor EE, Baldwin RL, Tenesa A, Ma L, et al. 2019. Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biol* **17:** 68. doi:10.1186/s12915-019-0687-8
Fang L, Cai W, Liu S, Canela-Xandri O, Gao Y, Jiang J, Rawlik K, Li B, Schroeder SG, Rosen BD, et al. 2020. Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. *Genome Res* **30:** 790–801. doi:10.1101/gr.250704.119
Fink T, Lopdell TJ, Tiplady K, Handley R, Johnson TJJ, Spelman RJ, Davis SR, Snell RG, Littlejohn MD. 2020. A new mechanism for a familiar mutation: Bovine DGAT1 K232A modulates gene expression through multi-junction exon splice enhancement. *BMC Genomics* **21:** 591. doi:10.1186/s12864-020-07004-z
Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, Esquerré D, Zytnicki M, Derrien T, Bardou P, et al. 2019. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol* **17:** 108. doi:10.1186/s12915-019-0726-5
Freking BA, Murphy SK, Wylie AA, Rhodes SJ, Keele JW, Leymaster KA, Jirtle RL, Smith TPL. 2002. Identification of the single base change causing the callipyge muscle hypertrophy phenotype, the only known example of polar overdominance in mammals. *Genome Res* **12:** 1496–1506. doi:10.1101/gr.571002
García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassell CP. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc Natl Acad Sci* **113:** E3995–E4004. doi:10.1073/pnas.1519061113
Georges M, Charlier C, Smit M, Davis E, Shay T, Tordoir X, Takeda H, Caiment F, Cockett N. 2004. Toward molecular understanding of polar overdominance at the ovine callipyge locus. *Cold Spring Harb Symp Quant Biol* **69:** 477–484. doi:10.1101/sqb.2004.69.477
Graf A, Krebs S, Zakhartchenko V, Schwalb B, Blum H, Wolf E. 2014. Fine mapping of genome activation in bovine embryos by RNA sequencing. *Proc Natl Acad Sci* **111:** 4139–4144. doi:10.1073/pnas.1321569111
Halstead MM, Kern C, Saelao P, Wang Y, Chanthavixay G, Medrano JF, Van Eenennaam AL, Korf I, Tuggle CK, Ernst CW, et al. 2020a. A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. *BMC Genomics* **21:** 698. doi:10.1186/s12864-020-07078-9
Halstead MM, Ma X, Zhou C, Schultz RM, Ross PJ. 2020b. Chromatin remodeling in bovine embryos indicates species-specific regulation of genome activation. *Nat Commun* **11:** 4654. doi:10.1038/s41467-020-18508-3
Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589. doi:10.1016/j.molcel.2010.05.004
Johnston D, Kim J, Taylor JF, Earley B, McCabe MS, Lemon K, Duffy C, McMenamy M, Cosby SL, Waters SM. 2021. ATAC-seq identifies regions of open chromatin in the bronchial lymph nodes of dairy calves experimentally challenged with bovine respiratory syncytial virus. *BMC Genomics* **22:** 14. doi:10.1186/s12864-020-07268-5
Kaiser VB, Talmane L, Kumar Y, Semple F, MacLennan M, Deciphering Developmental Disorders Study, FitzPatrick DR, Taylor MS, Semple CA. 2021. Mutational bias in spermatogonia impacts the anatomy of regulatory sites in the human genome. *Genome Res* **31:** 1994–2007. doi:10.1101/gr.275407.121
Karim L, Takeda H, Lin L, Druet T, Arias JAC, Baurain D, Cambisano N, Davis SR, Farnir F, Grisart B, et al. 2011. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nat Genet* **43:** 405–413. doi:10.1038/ng.814
Kern C, Wang Y, Xu X, Pan Z, Halstead M, Chanthavixay G, Saelao P, Waters S, Xiang R, Chamberlain A, et al. 2021. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nat Commun* **12:** 1821. doi:10.1038/s41467-021-22100-8
Khansefid M, Pryce JE, Bolormaa S, Chen Y, Millen CA, Chamberlain AJ, Vander Jagt CJ, Goddard ME. 2018. Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle. *BMC Genomics* **19:** 793. doi:10.1186/s12864-018-5181-0
Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12:** 357–360. doi:10.1038/nmeth.3317
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359. doi:10.1038/nmeth.1923
Lee Y-L, Takeda H, Costa Monteiro Moreira G, Karim L, Mullaart E, Coppieters W, The GplusE consortium, Appeltant R, Veerkamp RF, Groenen MAM, et al. 2021. A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy cattle. *PLoS Genet* **17:** e1009331. doi:10.1371/journal.pgen.1009331
Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478:** 476–482. doi:10.1038/nature10530
Liu X, Li YI, Pritchard JK. 2019. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177:** 1022–1034.e6. doi:10.1016/j.cell.2019.04.014
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/s13059-014-0550-8
Luquette LJ, Miller MB, Zhou Z, Bohrson CL, Zhao Y, Jin H, Gulhan D, Ganz J, Bizzotto S, Kirkham S, et al. 2022. Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. *Nat Genet* **54:** 1564–1571. doi:10.1038/s41588-022-01180-2
Markljung E, Jiang L, Jaffe JD, Mikkelsen TS, Wallerman O, Larhammar M, Zhang X, Wang L, Saenz-Vash V, Gnirke A, et al. 2009. ZBED6, a novel transcription factor derived from a domesticated DNA transposon regulates IGF2 expression and muscle growth. *PLoS Biol* **7:** e1000256. doi:10.1371/journal.pbio.1000256
Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Teodosiadis A, et al. 2020. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584:** 244–251. doi:10.1038/s41586-020-2559-3
Ming H, Sun J, Pasquariello R, Gatenby L, Herrick JR, Yuan Y, Pinto CR, Bondioli KR, Krisher RL, Jiang Z. 2021. The landscape of accessible chromatin in bovine oocytes and early embryos. *Epigenetics* **16:** 300–312. doi:10.1080/15592294.2020.1795602
Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M, Klein M, Hildebrandt J, Neumann M, Kliebenstein D, et al. 2022. Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature* **602:** 101–105. doi:10.1038/s41586-021-04269-6
Nielsen R, Slatkin M. 2013. *An introduction to population genetics: theory and applications*. Oxford University Press, Oxford, New York.
Oget-Ebrad C, Kadri NK, Moreira GCM, Karim L, Coppieters W, Georges M, Druet T. 2022. Benchmarking phasing software with a whole-genome sequenced cattle pedigree. *BMC Genomics* **23:** 130. doi:10.1186/s12864-022-08354-6
Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33:** 290–295. doi:10.1038/nbt.3122
Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv doi:10.1101/201178
Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44:** W160–W165. doi:10.1093/nar/gkw257
R Core Team. 2023. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.
Reijns MAM, Kemp H, Ding J, de Procé SM, Jackson AP, Taylor MS. 2015. Lagging-strand replication shapes the mutational landscape of the genome. *Nature* **518:** 502–506. doi:10.1038/nature14183
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26. doi:10.1038/nbt.1754
Sabarinathan R, Mularoni L, Deu-Pons J, Gonzales-Perez A, López-Bigas N. 2016. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532:** 264–267. doi:10.1038/nature17661
Smit M, Segers K, Carrascosa LG, Shay T, Baraldi F, Gyapay G, Snowder G, Georges M, Cockett N, Charlier C. 2003. Mosaicism of solid gold supports the causality of a noncoding A-to-G transition in the determinism of the callipyge phenotype. *Genetics* **163:** 453–456. doi:10.1093/genetics/163.1.453
Stegle O, Parts L, Durbin R, Winn J. 2010. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* **6:** e1000770. doi:10.1371/journal.pcbi.1000770
Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci* **100:** 9440–9445. doi:10.1073/pnas.1530509100
Trynka G, Westra H-J, Slowikowski K, Hu X, Xu H, Stranger BE, Klein RJ, Han B, Raychaudhuri S. 2015. Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *Am J Hum Genet* **97:** 139–152. doi:10.1016/j.ajhg.2015.05.016

Van Laere A-S, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M, et al. 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425:** 832–836. doi:10.1038/nature02064

Wathes DC, Becker F, Buggiotti L, Crowe MA, Ferris C, Foldager L, Grelet C, Hostens M, Ingvartsen KL, Marchitelli C, et al. 2021a. Associations between circulating IGF-1 concentrations, disease status and the leukocyte transcriptome in early lactation dairy cows. *Ruminants* **1:** 147–177. doi:10.3390/ruminants1020012

Wathes DC, Cheng Z, Salavati M, Buggiotti L, Takeda H, Tang L, Becker F, Ingvartsen KI, Ferris C, Hostens M, et al. 2021b. Relationships between metabolic profiles and gene expression in liver and leukocytes of dairy cows in early lactation. *J Dairy Sci* **104:** 3596–3616. doi:10.3168/jds.2020-19165

Xiang R, van den Berg I, MacLeod IM, Hayes BJ, Prowse-Wilkins CP, Wang M, Bolormaa S, Liu Z, Rochfort SJ, Reich CM, et al. 2019. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc Natl Acad Sci* **116:** 19398–19408. doi:10.1073/pnas.1904159116

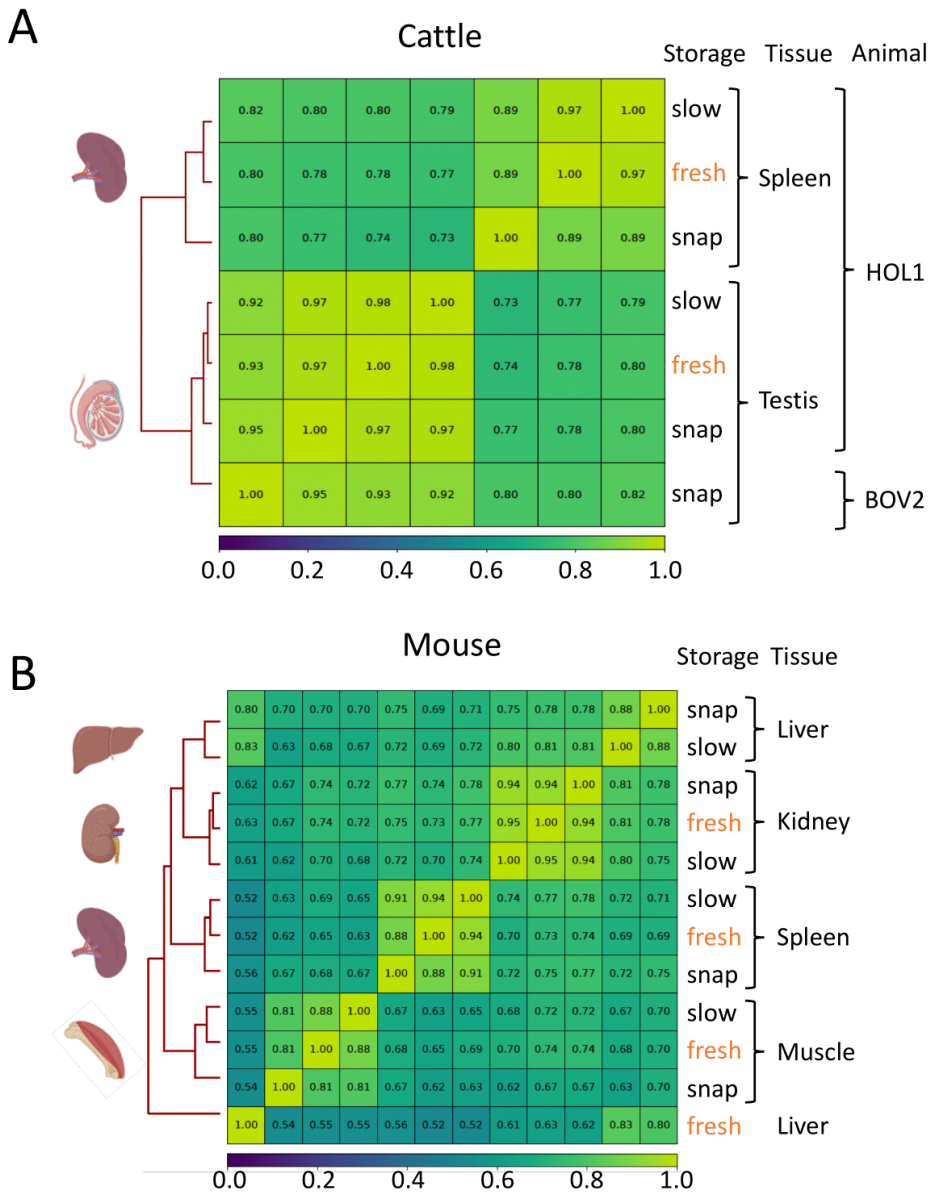
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-seq (MACS). *Genome Biol* **9:** R137. doi:10.1186/gb-2008-9-9-r137

Received April 1, 2023; accepted in revised form September 19, 2023.

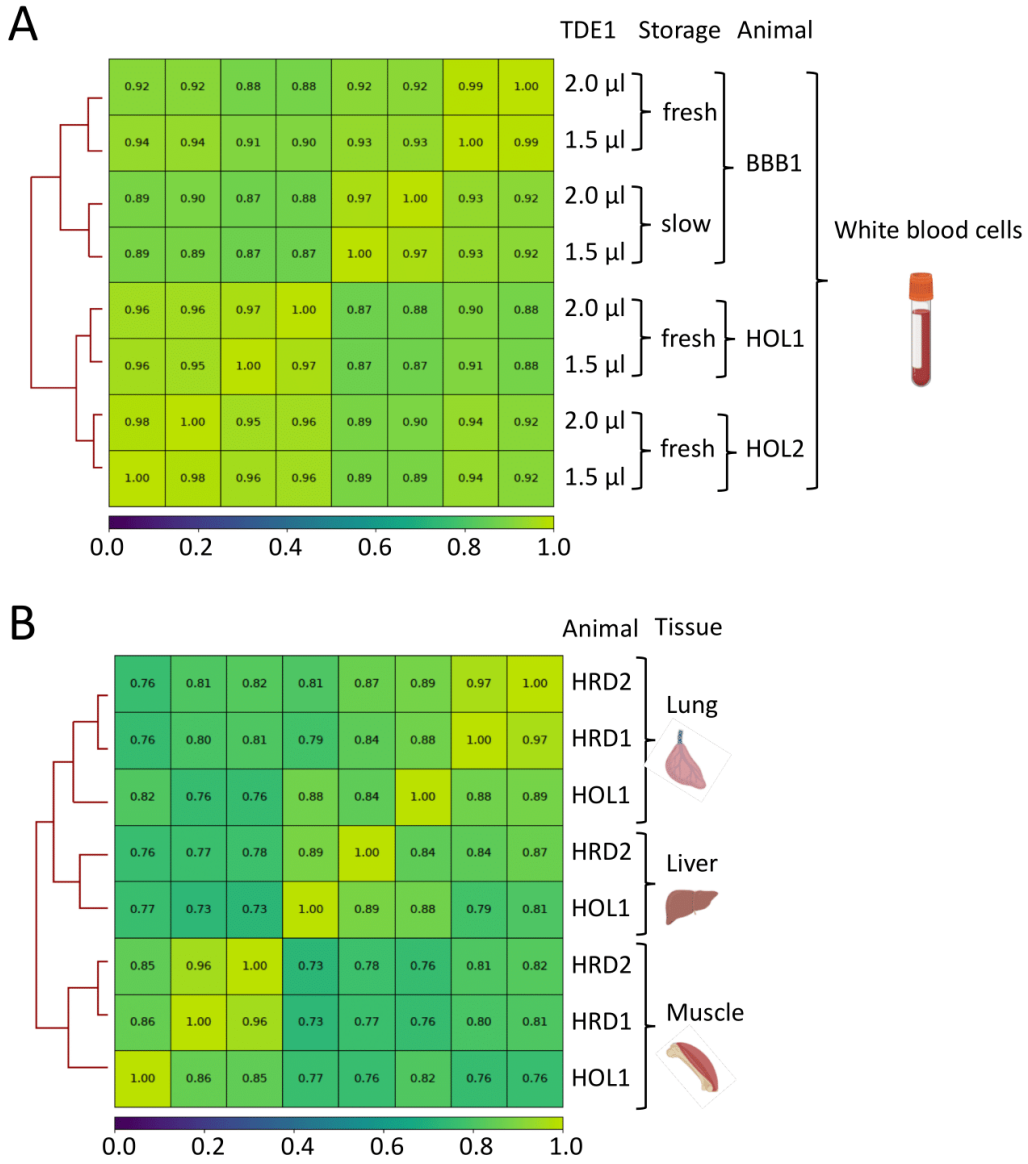
An organism-wide ATAC-seq peak catalog for the bovine and its use to identify regulatory variants

Can Yuan, Lijing Tang, Thomas Lopdell, et al.

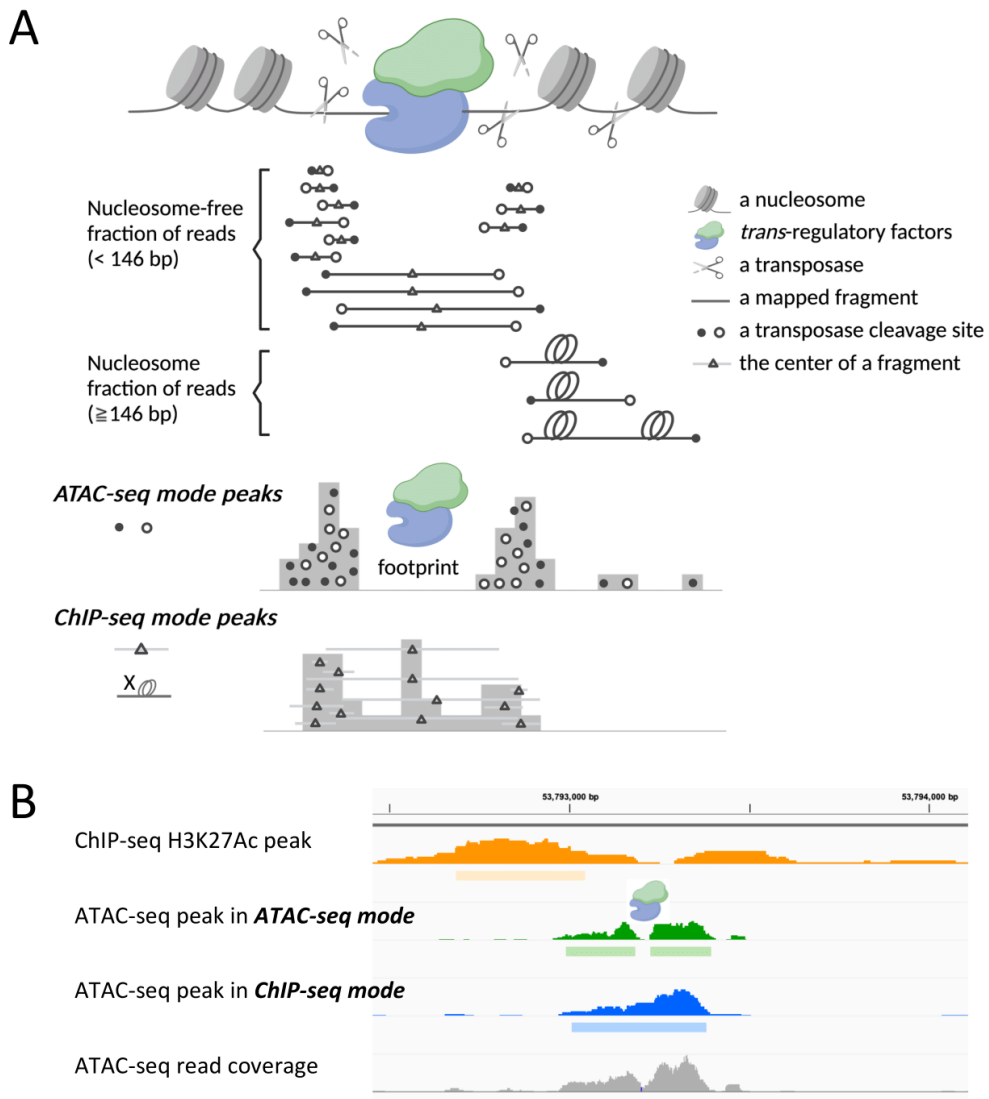
Genome Res. 2023 33: 1848-1864 originally published online September 26, 2023
Access the most recent version at doi:10.1101/gr.277947.123



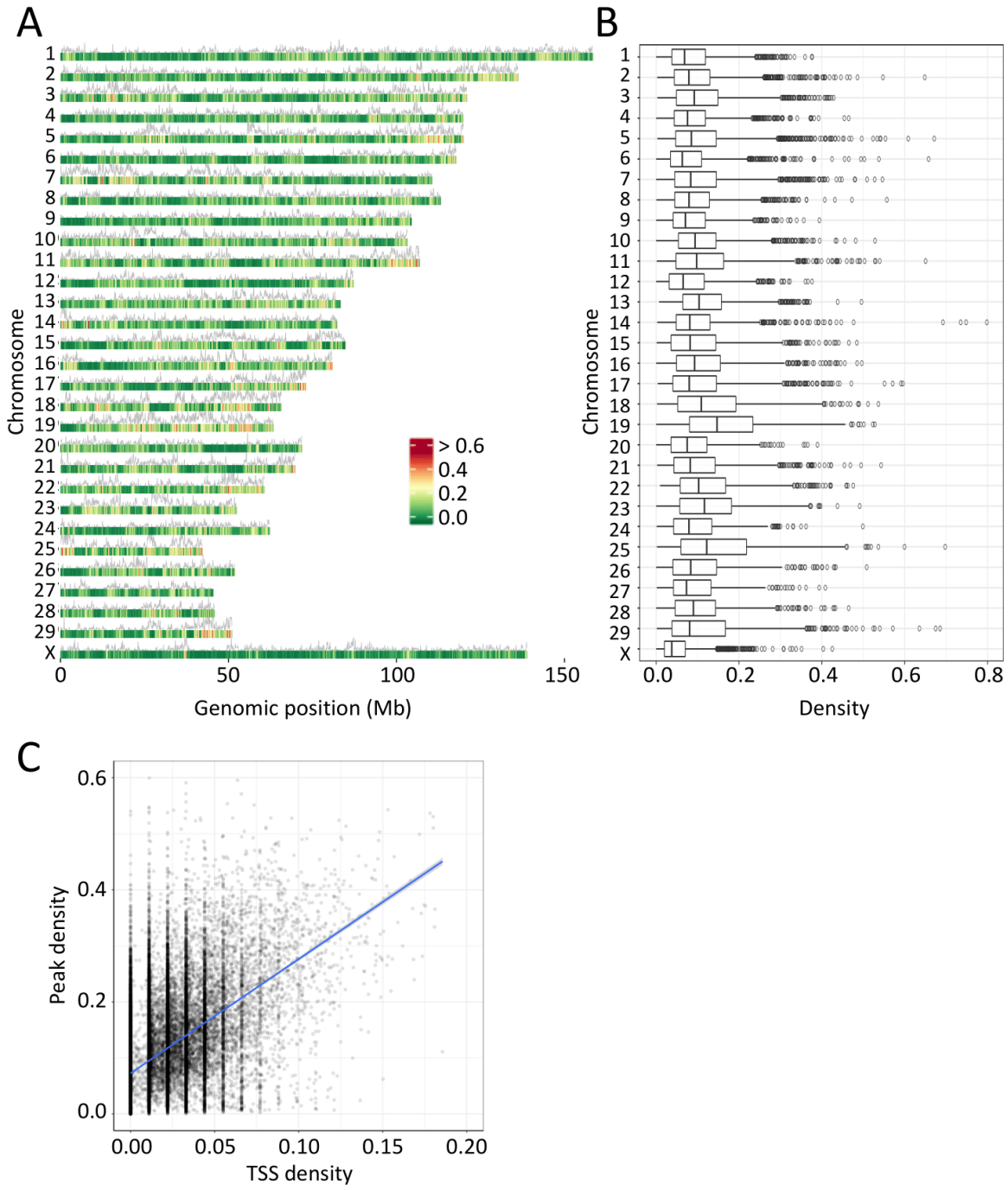
Supplemental Figure S1: Impacts of sample storage conditions on genome-wide ATAC-seq signals were examined using cattle **(A)** and mouse **(B)** tissues. Pearson's correlation was calculated using normalized read coverage in 500-bp windows spanning the entire genome. Pairwise correlation coefficients are shown in color code between hierarchically clustered samples. Tissues were collected in the UW cold storage solution on ice and further processed in three ways: (i) fresh: Tissues were directly subjected to library construction on the day of sampling, (ii) slow: Tissues cut into ~ 3 mm cubes were soaked in STEM-CELLBANKER DMSO-free cell freezing medium for ~ 10 minutes on ice and stored at -80°C until use, and (iii) snap: Tissues cut into ~ 3 mm cubes were snap-frozen in liquid nitrogen and stored at -80°C until use. Correlations between fresh and slow-frozen samples are slightly higher than between fresh and snap-frozen samples. This suggests that cell freezing medium better preserves nuclear structure of cryopreserved tissues. It is noteworthy that, for tissues that are difficult to homogenize (e.g., stomach, nerve, vein), pulverizing snap-frozen tissues using mortar and pestle in liquid nitrogen achieved better results (see also Supplemental Table S3).



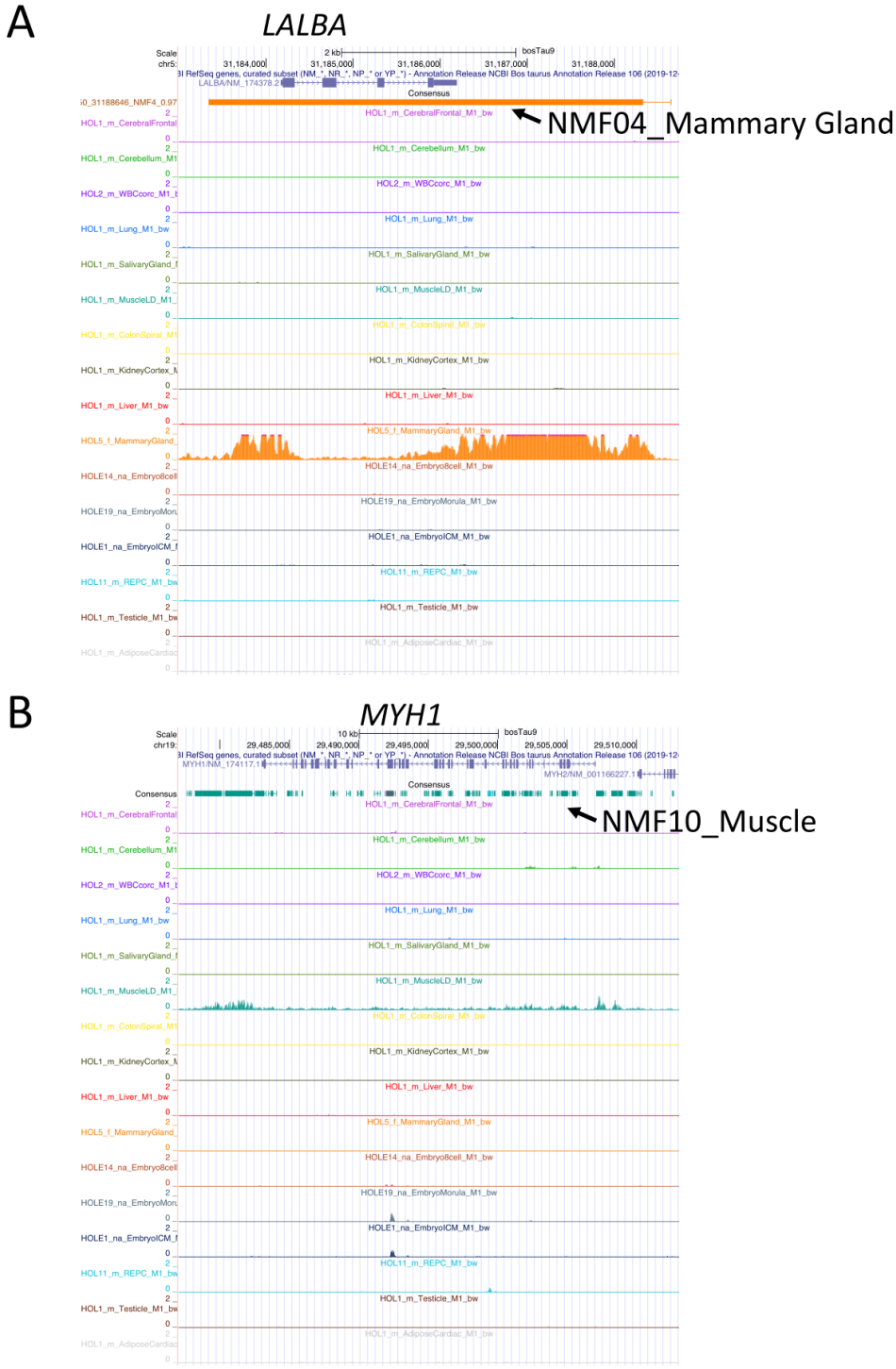
Supplemental Figure S2: Reproducibility of ATAC-seq data. Pairwise correlation coefficients calculated using normalized read coverage in 500-bp windows are shown in color code for hierarchically clustered samples. **(A)** Comparisons between in-house ATAC-seq data using white blood cells (WBC) isolated from one Belgium Blue adult female (BBB1) and two Holstein juvenile male (HOL1 and HOL2). Samples were either directly subjected to library construction (fresh) or stored in the Stem-Cellbanker freezing media until use (slow). Tagmentation reactions were performed using two different amounts of TDE1 transposase enzyme (1.5 µl or 2.0 µl) per 50,000 nuclei. **(B)** Correlation between in-house (HOL1) and publicly available ATAC-Seq data (HRD1 and HRD2, two adult Hereford cattle, Halstead *et al.*, 2020). Genome-wide ATAC-Seq signal was highly reproducible between technical (0.92 – 0.99) as well as biological replicates (0.85 – 0.97).



Supplemental Figure S3: (A) ATAC-seq peaks were called in two ways: (i) ATAC-seq mode focusing on open chromatin regions. Genomic loci cleaved by transposase (tags) (shown in small circles) were defined as 38-bp regions around the reads' 5'-ends (Adey *et al.*, 2010, Buenrostro *et al.*, 2013). Peaks (open chromatin regions) were identified by comparing the tag distribution of a sample to one from a purified genomic DNA control; and (ii) ChIP-seq mode to cover regions protected from transposase cleavage events due to binding(s) of *trans*-regulatory factor(s) like a transcription factor (so-called footprints in the ATAC-seq mode analysis). Peaks were called with a setting generally used for ChIP-seq analysis (MACS2 piles up entire sequencing fragments instead of focusing on transposase cleavage sites close to 5'-ends of reads). To avoid covering nucleosome positions, only nucleosome-free fraction of reads (mapped fragment size < 146 bp) was used. This figure was created with BioRender.com. **(B)** An example of liver ATAC-seq peaks called in either ATAC-seq (shown in green) or ChIP-seq (blue) mode in a putative enhancer region (chr3:53,792,472-53,793,916) overlapping with H3K27Ac ChIP-seq peaks in a cognate liver tissue (orange) (Villar *et al.*, 2015). ChIP-seq mode of peaks facilitate to cover and identify *trans*-acting factor's footprints depleted in ATAC-seq mode peak calling.

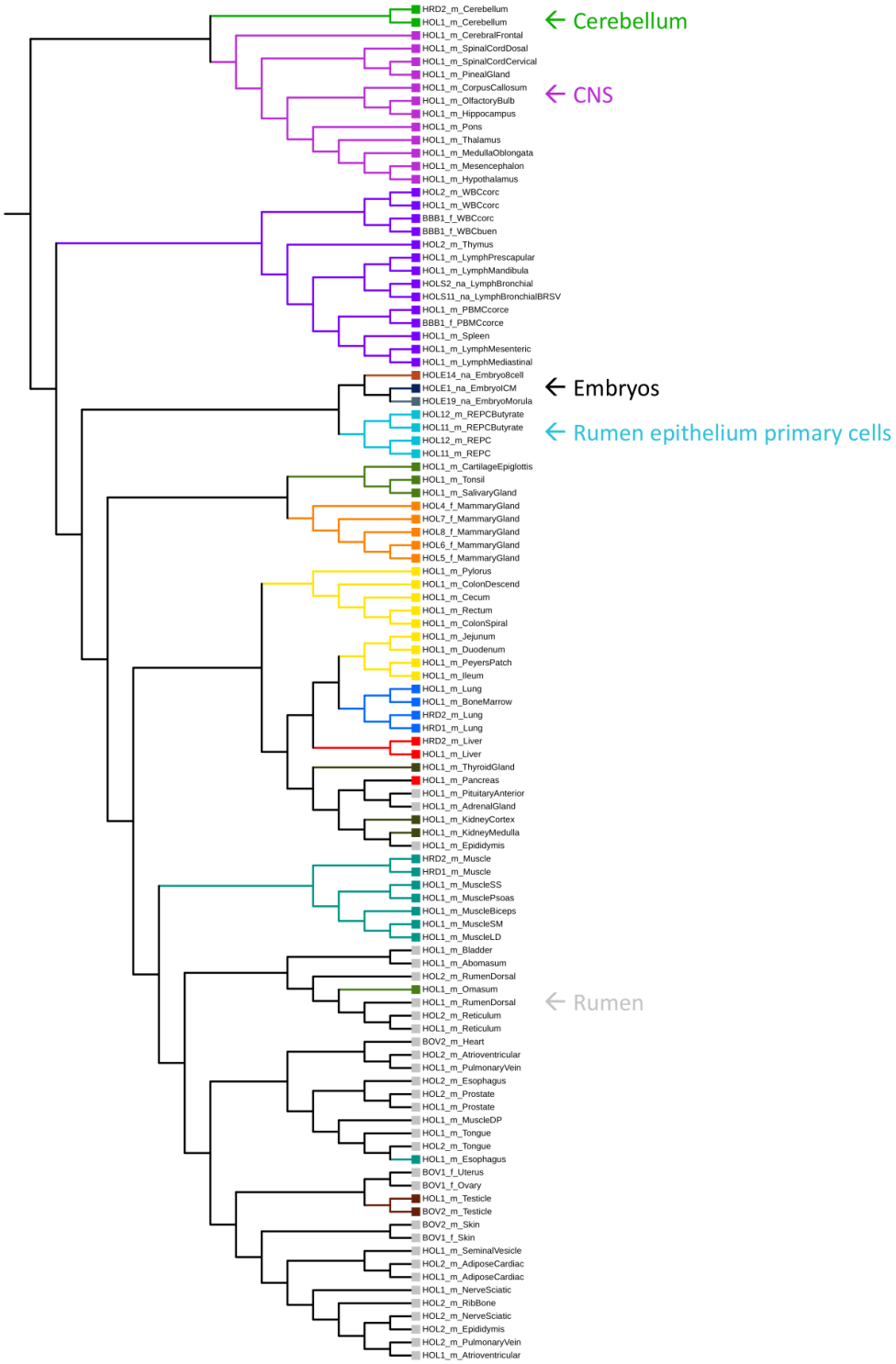


Supplemental Figure S4: (A) Density of ATAC-seq peaks (proportion of genome space occupied by ATAC-seq peaks) in 0.1 Mb sliding windows across the bovine genome in pseudo-color (green: low – red: high). Grey curves: density of TSS (number of TSS) in the same 0.1 Mb sliding windows. **(B)** Distribution of density of ATAC-seq peaks in 0.1 Mb sliding windows by chromosome. **(C)** Correlation between TSS (x-axis) and peak (y-axis) density in 0.1 Mb windows (r = 0.52).

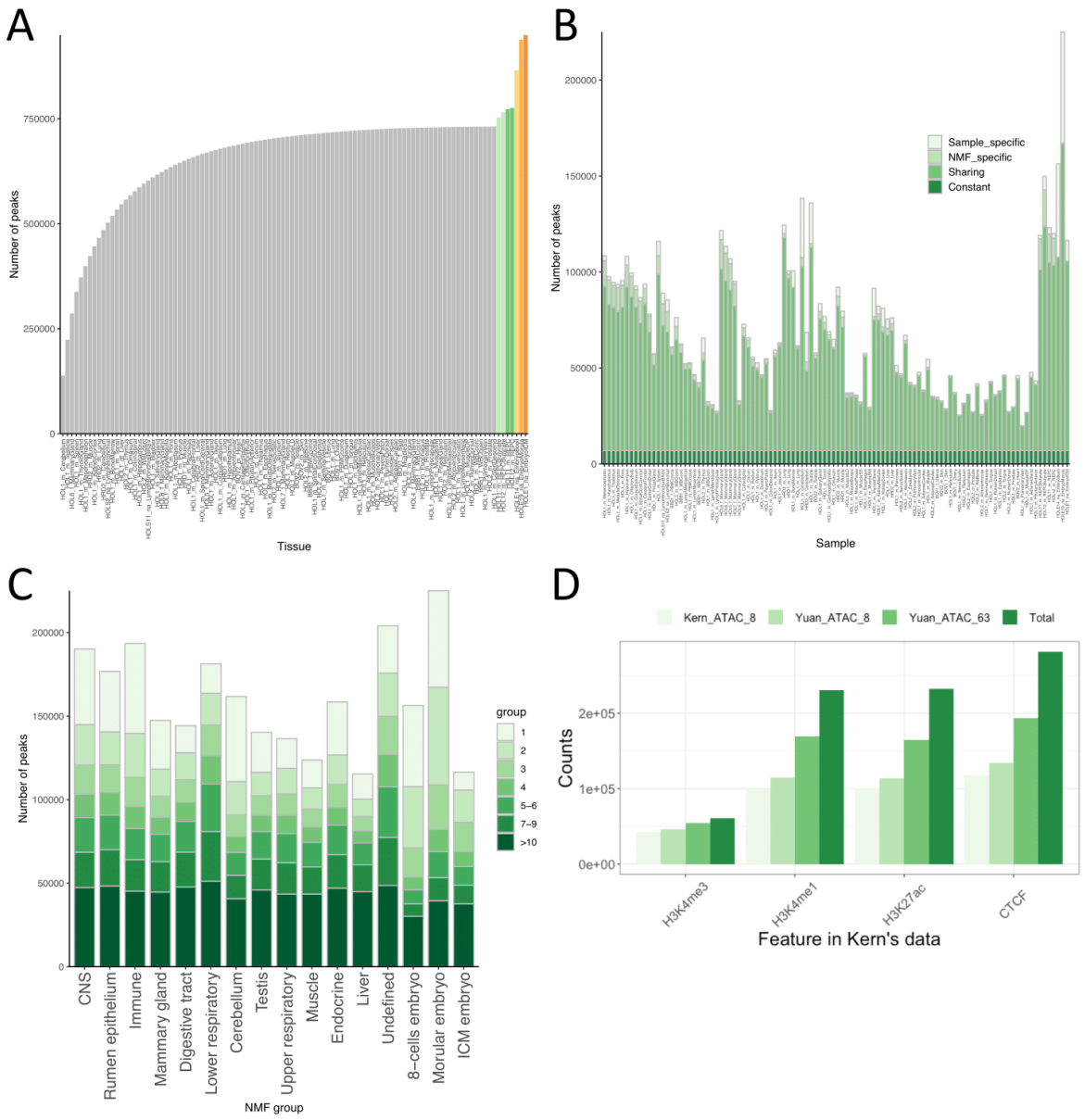


Supplemental Figure S5: UCSC Genome Browser view of tissue-type-specific ATAC-seq peaks and their NMF annotations (https://www.gigauag.uliege.be/cms/c_4791343/en/gigauag-diagnostics-software-data) around **(A)** mammary gland specific gene, *LALBA* (*Lactalbumin Alpha*) and **(B)** muscle specific gene, *MYH1* (*Myosin Heavy Chain 1*) where chromatin is opened in a tissue specific way across the genes.

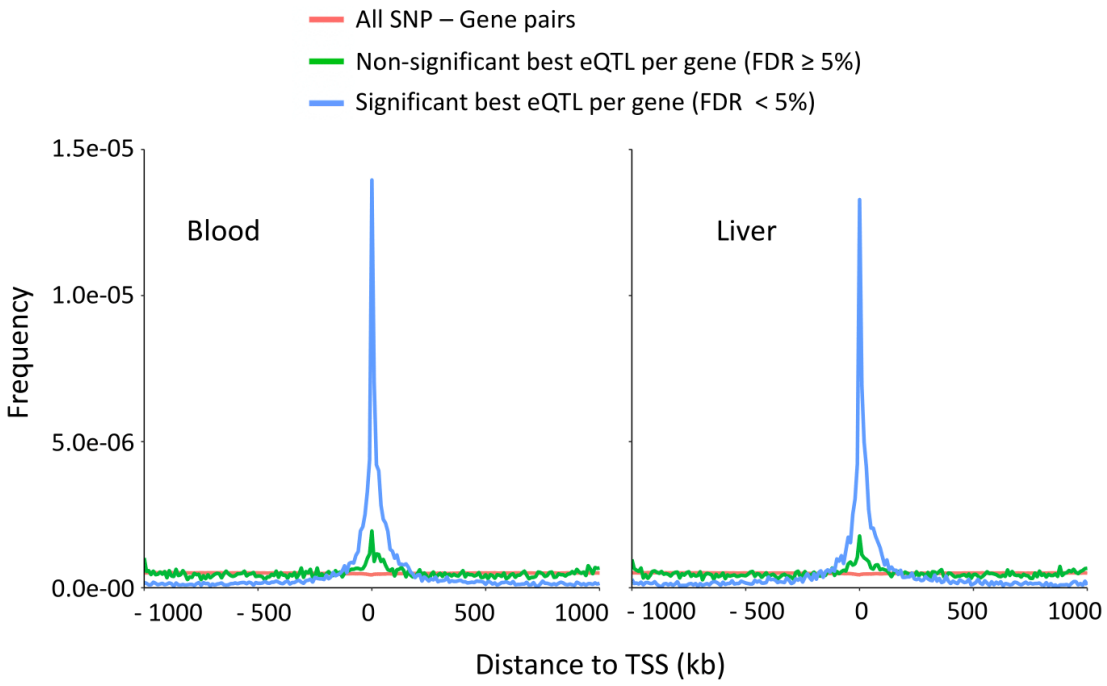
Supplemental Figure S6: UCSC Genome Browser view of ATAC-seq peaks and their NMF annotations (https://www.gigauag.uliege.be/cms/c_4791343/en/gigauag-diagnostics-software-data) around *PLCD4* (*Phospholipase C Delta 4*) gene. A recessive mutation in this gene has pleiotropic effects on multiple cow phenotypes (Reynolds *et al.*, 2021). The gene might be regulated by various combinations of shared and tissue-type-specific regulatory elements. Peak names consist of chromosome coordinate, largest NMF component and its weight. Peaks are color-coded as in the main Fig. 1A and F.



Supplemental Figure S7: Hierarchical clustering of the 104 samples using the Ward.D2 method implemented with the hclust R function, and the accessibility (measured as the fold-enrichment over genomic DNA background at the nucleotide position in the peak with the highest such value) of 934,972 distal (and hence more tissue specific) peaks. The samples are color-coded by their highest NMF component as in main Fig. 1A and F. Samples discussed in the main text are marked.



Supplemental Figure S8: (A) Number of additional ATAC-seq peaks that are uncovered when adding new samples. The first 97 samples are the post-natal tissue samples, ordered by decreasing number of new peaks detected. Primary cultured cells (green gradients) and embryonic samples (orange gradient) are added at the end. **(B)** Number of peaks detected in each sample colored by degree of sharing with other samples (Constant: shared with all other samples, Sharing: shared with at least one sample from other NMF component(s), NMF_specific: shared only with samples from the same NMF component, Sample_specific: not shared with any other sample). **(C)** Number of peaks sorted by NMF component (dominant component) and colored by the number of NMF components with whom it is shared. **(D)** Overlap between ATAC-seq and ChIP-seq epigenetic marks in the dataset of Kern *et al*. (2021) studying eight tissue types in cattle. H3K4me3: active promotors, H3K4me1: active and primed enhancers, H3K27ac: active promotors and enhancers, CTCF: boundary elements and regulators of transcription and chromatin architecture. H3K27me3, corresponding to silenced chromatin, was not included (although studied in Kern *et al*.) because they do not correspond to regulatory elements *sensu stricto* but rather reflect the local outcome of gene repression. Kern_ATAC_8: ATAC-seq peaks reported in Kern *et al.* from the analyses of eight tissue types. Yuan_ATAC_8: ATAC-seq peaks detected in this study from the analyses of eight tissue-types matching the Kern *et al*. samples. Yuan_ATAC_63: full ATAC-seq peak catalogue produced in this study.



Supplemental Figure S9: Frequency distribution of the positions, with respect to the TSSs, of the lead eQTL SNVs for significant eQTL (blue), the lead eQTL SNVs for non-significant eQTL (green), and all SNVs.

Experimental Section

Study 2

Evaluation of heritability partitioning approaches in livestock populations

BMC Genomics 25(1):690

Can Yuan, José Luis Gualdrón Duarte, Haruko Takeda, Michel Georges and Tom Druet

4 Experimental section: Study 2

4.1 Summary

Heritability partitioning approaches estimate the contribution of different functional classes, such as coding or regulatory variants, to the genetic variance. This information allows a better understanding of the genetic architecture of complex traits, including complex diseases, but can also help improve the accuracy of genomic selection in livestock species. However, methods have mainly been tested on human genomic data, whereas livestock populations have specific characteristics, such as high levels of relatedness, small effective population size or long-range levels of linkage disequilibrium. Here, we used data from 14,762 cows, imputed at the whole-genome sequence level for 11,537,240 variants, to simulate traits in a typical livestock population and evaluate the accuracy of two state-of-the-art heritability partitioning methods, GREML and a Bayesian mixture model. In simulations where a single functional class had increased contribution to heritability, we observed that the estimators were unbiased but had low precision. When causal variants were enriched in variants with low (<0.05) or high (> 0.20) minor allele frequency or low (below 1st quartile) or high (above 3rd quartile) linkage disequilibrium scores, it was necessary to partition the genetic variance into multiple classes defined on the basis of allele frequencies or LD scores to obtain unbiased results. When multiple functional classes had variable contributions to heritability, estimators showed higher levels of variation and confounding between certain categories was observed. In addition, estimators from small categories were particularly imprecise. However, the estimates and their ranking were still informative about the contribution of the classes. We also demonstrated that using methods that estimate the contribution of a single category at a time, a commonly used approach, results in an overestimation. Finally, we applied the methods to phenotypes for muscular development and height and estimated that, on average, variants in open chromatin regions had a higher contribution to the genetic variance (> 45%), while variants in coding regions had the strongest individual effects (> 25-fold enrichment on average). Conversely, variants in intergenic or intronic regions showed lower levels of enrichment (0.2 and 0.6-fold on average, respectively). Heritability partitioning approaches should be used cautiously in livestock populations, in particular for small categories. Two-component approaches that fit only one functional category at a time lead to biased estimators and should not be used.

4.2 Introduction

In livestock species, the number of genotyped and whole-genome sequenced animals is steadily increasing. Combining these data with missing genotype imputation techniques allows genome-wide association studies and genomic selection to be performed at the sequence level in large cohorts. More recently, functional annotations of the genome are becoming available for several livestock species (Andersson et al., 2015; Clark et al., 2020). For example, transcriptome data (Fang et al., 2020; Liu et al., 2022), chromatin accessibility maps (Kern et al., 2021; Yuan et al., 2023) or histone mark distributions (Kern et al., 2021; S. Liu et al., 2020) are now available in cattle. In human genetics, such information has been used to study the genetic architecture of complex traits, including complex diseases (Finucane et al., 2015; Gusev et al., 2014). More precisely, the contribution of different functional categories of variants to the genetic variance of these different traits has been estimated. Such approaches are referred to as variance partitioning or heritability partitioning approaches. They have for example highlighted the importance of regulatory variants (Finucane et al., 2015; Gusev et al., 2014). Fewer studies have been realized in livestock species, as functional annotation maps remain limited compared to humans, and are more recent. Nevertheless, similar approaches have been used, for instance, in cattle (Koufariotis et al., 2014; Xiang et al., 2019a). In this context, the identification of functional categories contributing to complex traits is also important for prioritizing variants to be used in genomic selection and improving its accuracy.

Most methods used for heritability partitioning have been developed and tested in the context of human genetics (Finucane et al., 2015; Patxot et al., 2021; Yang et al., 2015). Although livestock species have specific characteristics at the genomic level, methods have often been transferred without additional testing. As a result of their demographic history, including domestication, breed creation and intensive selection, livestock species are indeed different in terms of effective population size (Hayes et al., 2003; MacLeod et al., 2013), levels and extent of linkage disequilibrium (LD) (Gautier et al., 2007), relatedness between individuals and levels of inbreeding (Leroy, 2014). The higher selection intensity in livestock species often results in the fixation of large effect variants accompanied by large selective sweeps (Druet et al., 2013). Importantly, previous studies in humans have relied on samples of unrelated individuals, discarding all pairs of individuals with a relatedness level above 0.025 (Yang et al., 2015), whereas in a typical livestock dataset these and higher relationships are common. For instance, with the use of artificial insemination, many individuals may have a common sire or grand-sire. Similarly, the importance of accounting for LD scores when estimating variance components (Yang et al., 2015) has not been evaluated when high LD levels are present at long distances (Farnir et al., 2000).

We herein used a genotyped population of 14,762 Belgian Blue Beef (BBB) cows to evaluate the accuracy of heritability partitioning approaches in a typical livestock population. Belgian Blue cattle have indeed been intensively selected for muscular development. This has resulted in the fixation of an 11bp deletion in the myostatin gene (Grobet et al., 1997), accompanied by a large selective sweep (Druet

et al., 2013). Additional genetic variation for muscular development has been exploited to further improve this trait (Druet et al., 2014a). As in other livestock populations, the effective population size is small, around 100 (Druet et al., 2013), and individuals have high levels of recent inbreeding associated with long runs of homozygosity (Solé et al., 2017). The objective of the present study was to use these data to perform realistic simulations, with characteristics of a typical livestock population, in order to evaluate two state-of-the-art methods, a variance component approach (Yang et al., 2015) and a Bayesian mixture model (Patxot et al., 2021). The simulations included scenarios where causal variants were enriched in specific allele frequency, LD score or functional categories. An additional objective was to use these approaches to perform heritability partitioning based on functional annotation for muscular development and height traits in Belgian Blue beef cattle.

4.3 Material and methods

4.3.1 Data

For the present study we used data from 14,762 Belgian Blue beef cows with imputed genotypes from 11,537,240 SNPs and small indels (Gualdrón Duarte et al., 2023). Cows were genotyped with either low-density (9983 to 20,502 SNPs) or medium-density (51,809 to 57,979 SNPs) arrays and genotype imputation to the sequence level was performed in successive steps. The reference panels included 13,600, 890 and 230 individuals at the medium-density (28,893 SNPs selected), high-density (572,667 SNPs selected) and sequence levels, respectively. Variants with low minor allele frequency (MAF) (< 0.01) or with lower imputation accuracy ($r^2 < 0.90$) were filtered out, resulting in the selection of 11,431,742 variants. More details on the imputation procedure and the data set can be found in Gualdron Duarte et al. (2023). We used phenotypes for muscularity traits (shoulder muscularity, top muscularity, buttock muscularity rear and side view) and height (with heritabilities of 0.30, 0.31, 0.42, 0.39 and 0.38, respectively). The four muscularity traits are scores from 51 to 100, given on the farm by a technician based on a visual assessment (available for 14,476 individuals), while height was measured for 12,904 individuals. In addition, a synthetic score for muscular development was obtained as a linear combination of the four individual muscularity scores (with a weight of 1 for shoulder and top muscularity and 2 for buttock muscularity scores). These phenotypes were corrected for fixed effects from the evaluation model as described in Gualdron Duarte et al. (2023).

4.3.2 Variant annotation

For variant annotation, we selected categories similar to those defined by Gusev et al. (2014). Accordingly, six functional categories were defined to classify the 11,431,742 variants. First, we identified variants located in open chromatin regions (OCR). These regions were defined using an organism-wide catalog of 976,813 cis-acting regulatory elements for the bovine detected by the assay for transposase accessible chromatin using sequencing (ATAC-Seq) described in Yuan et al. (2023).

The catalogue was generated using data from 106 samples corresponding to 68 tissue types. We annotated as OCR variants those variants located in the 976,813 peaks, which represented 10% of the genome space. Variants outside the OCR were classified into five additional groups corresponding to coding sequence (CDS), untranslated regions (UTR) including both 3' and 5' UTR, regions upstream (-1kb) or downstream (+1kb) of genes (UDR), intronic (IOR) and intergenic (IGR) regions. The number of variants per category is reported in Table1. This annotation was obtained from the General Transfer Format (GTF) file of the bovine genome assembly ARS-UCD1.2 downloaded from Ensembl (v105). This file directly provides coordinates of genes, transcripts, exons, CDS and UTR. IORs were defined as non-exonic regions in genes. Transcription start and termination sites (TSS and TTS) were obtained using Homer (Heinz et al., 2010) and all transcripts from the genes. Upstream and downstream regions were then defined as 1 kb upstream and downstream from TSS and TTS, respectively. IGR corresponded to the remaining unannotated regions.

Annotation groups were also defined based on MAF and linkage disequilibrium (LD) scores (Yang et al., 2015). Three MAF groups were defined [0.01-0.05; 0.05-0.10; 0.10-0.50]. For each variant, LD scores were obtained using GCTA (Yang et al., 2011a) as the sum of LD r^2 scores between the variant and all variants within a 200 kb window (Yang et al., 2015). SNPs were then stratified into four LD score groups based on quartiles. These groups thus represent SNPs that have, for example, low or high LD levels with other SNPs in the region. SNPs in high LD groups capture the effect of more SNPs, and potentially causal variants, than SNPs in low LD groups.

4.3.3 Heritability partitioning methods

Two methods were applied to estimate the contribution of different annotation groups to the additive genetic variance. First, we used a genomic restricted maximum likelihood (GREML) approach to estimate the variance components with the following linear mixed model:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \sum_{s=1}^{S} \boldsymbol{g}_s + \boldsymbol{e},$$

where y is the vector of individual phenotypes, $\mathbf{1}\boldsymbol{\mu}$ is the intercept term (i.e. the mean effect), g_s is the vector of individual polygenic effects associated to annotation group s, S is the total number of fitted annotation groups, and e is the vector of individual random error terms. Each polygenic component is normally distributed, $\boldsymbol{g}_s \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{G}_s\sigma_s^2)$ where $\boldsymbol{G}_s$ is the genomic relationship matrix (GRM) computed using the variants present in category s and σ_s^2 is the variance of polygenic effects from the annotation group. The GRM were computed with GCTA using centered and scaled genotypes as described in Yang et al. (Yang et al., 2011a). The residual error terms are independent and normally distributed, $\boldsymbol{e} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{I}\sigma_e^2)$ where I is the identity matrix and σ_e^2 is the residual variance. The additive polygenic variance, σ_g^2, is equal to the sum of the variances associated to each annotation groups:

$$\sigma_g^2 = \sum_{s=1}^{S} \sigma_s^2.$$

The contribution of annotation group *s* to the genetic variance, called %SNP heritability, is estimated as:

$$\%h_s^2 = \frac{\sigma_s^2}{\sigma_g^2}.$$

Variance components were estimated using GCTA and the Average-Information (AI) algorithm (default option). When the AI-REML did not converge, we used the EM-REML algorithm with a maximum of 500 iterations. We follow Gusev et al. (Gusev et al., 2014) to define the enrichment of heritability as the percentage of heritability in category *s* divided by the proportion of variants in the same category.

The second approach is a Bayesian model designed for large-scale genomic data and called BayesRR-RC (Patxot et al., 2021). The model is an extension of BayesR (Erbe et al., 2012) and BayesRC (MacLeod et al., 2016). Variant effects are described as a mixture of null effects (spike probability at zero) and Gaussian distributions. The hyper-parameters vary for variants from different annotation groups. Accordingly, the variance explained by the markers and their mixture proportions are group-specific. Phenotypes are modeled as:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \sum_{s=1}^{S} \boldsymbol{X}_s \boldsymbol{\beta}_s + \boldsymbol{e},$$

where $\boldsymbol{X}_s$ is the matrix of centered and scaled genotypes for markers in category s and $\boldsymbol{\beta}_s$ is the vector of marker effect for category *s*. These effects are distributed according to:

$$\beta_{s_j} \sim \pi_{0_s}\delta_0 + \pi_{1_s}N(0, \sigma_{1_s}^2) + \pi_{2_s}N(0, \sigma_{2_s}^2) + \ldots + \pi_{L_s}N(0, \sigma_{L_s}^2),$$

where *j* is the marker index, δ_0 is a discrete probability mass at 0, *L* is the number of Gaussian distributions in the mixture, $\{\pi_{0_s}, \pi_{1_s}, \pi_{2_s}, \ldots, \pi_{L_s}\}$ are the mixture proportions for annotation group *s*, $\{\sigma_{1_s}^2, \sigma_{2_s}^2, \ldots, \sigma_{L_s}^2\}$ are the mixture variances for group *s*, proportional to σ_s^2, the variance explained by the group which is directly estimated from the data. In our study, we set *L* to 3, with variances $\sigma_{l_s}^2$ respectively equal to 0.0001, 0.001 and 0.01 σ_s^2. This model was run using the GMRM software (Patxot et al., 2021) with a Gibbs sampling scheme for 5,000 iterations with a burn-in period of 2,000 iterations. This setting corresponds to the values used by the software developers in their original study (Patxot et

al., 2021), and Orliac et al. (2022) have shown that 2,000 iterations allow to obtain good approximations of the parameters.

Different definitions of annotation groups can be applied in both approaches. In two-component (TC) models, two functional annotation groups are selected (e.g., OCR versus non-OCR), whereas in multiple-component (MC) models, multiple functional annotation groups are fitted simultaneously. Additional stratification levels can be added to these models (Yang et al., 2015). In the MAF-stratified (MS) and LD-stratified (LDS) models, groups are defined as a function of the MAF and LD score categories described above, respectively, whereas an LDMS model fits all combinations of functional, MAF and LD categories. In this case, the total number of fitted components is equal to the number of functional categories multiplied by the number of MAF groups and by the number of LD score groups. When a model is run without correcting for MAF or LD score categories, we use the abbreviation "noLDMS" to distinguish it from the other models.

4.3.4 Simulation study

Phenotype simulation. To obtain phenotypes with different architectures, we simulated them as:

$$\boldsymbol{y} = \sum_{s=1}^{S} \sum_{j=1}^{M_s} \boldsymbol{x}_{s_j} \beta_{s_j} + \boldsymbol{e}.$$

where y is the vector of individual simulated phenotypes, S is the number of different annotation groups, M_s is the number of causal variants (CVs) in annotation group s, $\boldsymbol{x}_{sj}$ is a vector of centered individual allele dosages for the j^{th} variant from the s^{th} group, β_{sj} is the effect of the corresponding variant and e is a vector of individual errors terms. By default, CV effect sizes were sampled from normal distributions with variance equal to $[2p_j(1\text{-}p_j)]^{-1}$, where p_j is the allele frequency of variant j. This is equivalent to assuming that each CV contributes equally to the genetic variance, as in Gusev et al. (2014) and Yang et al. (2015). This corresponds also to the default rule used by GCTA to construct the GRM. We assessed the robustness to this assumption later (see below). To simulate variable contributions of the annotation groups to the genetic variance, we selected the number of CVs, M_s, proportionally to the simulated contribution.

In this model, the individual polygenic effects g_i are equal to:

$$\boldsymbol{g} = \sum_{s=1}^{S} \sum_{j=1}^{M_s} \boldsymbol{x}_{s_j} \beta_{s_j} = \sum_{s=1}^{S} \boldsymbol{g}_s,$$

where g is the vector of individual polygenic effects and g_s is the vector of individual polygenic effects associated to annotation group s. After simulating these polygenic terms, their variance was rescaled to obtain the simulated contribution to the genetic variance, also defined as %SNP heritability. Finally, the individual error terms were normally distributed with a variance adjusted to obtain the simulated

heritability. By default, *M*, the total number of CVs, was set equal to 10,000 and the heritability to 0.50. This simulation code is available at https://github.com/can11sichuan/Bov-hg/.

Simulations scenarios with causal variants enriched in OCR. In unstratified scenarios, CVs were randomly sampled. In other scenarios, higher proportions of variants were sampled in certain annotation groups.

We started with simulations in which OCR contributed to 50% of the heritability, without stratification according to MAF or LD scores. Accordingly, 5,000 CVs were selected within OCR and 5,000 outside OCR. We then ran simulations in which CVs were enriched in specific MAF classes, LD-score categories, or combinations of both (LDMS simulation scenarios). The enriched annotation groups were defined as low MAF (MAF $\leq$ 0.05), high MAF (MAF > 0.20), low LD (LD scores below the 1st quartile) and high LD (LD scores above the 3rd quartile). In these simulations, 3,000 OCR SNPs were sampled in the enriched annotation groups and 2,000 OCR SNPs were sampled outside of these groups, and the same sampling was applied outside of OCR. A total of six stratified scenarios were defined: 1) low MAF, 2) high MAF, 3) low LD, 4) high LD, 5) low MAF and low LD, and 6) low MAF and high LD.

Finally, we tested the robustness of the approaches to the relationship between SNP effects and their MAF. In the default scenario described above, CVs have the same contribution to the genetic variance (i.e. rare variants have larger effects). In the alternative scenario corresponding to the first rules proposed by VanRaden (2008), the distribution of CV effects was independent of MAF (common variants would have a higher contribution to genetic variance).

Due to the high computational demands of BayesRR-RC, we worked with a subset of the genome. To do this, we randomly sampled 200 positions in the genome and selected all variants within 500 kb of the position (we sampled fragments rather than variants to preserve some LD structure). This resulted in a selection of 191 Mb and 965,428 variants (we have less than 200 Mb because some positions were less than 500 kb apart and their windows overlapped, while other positions were close to the chromosome ends). Both BayesRR-RC and GREML were applied to these simulations to ensure fair comparisons.

In total, each simulation scenario was repeated 100 times.

Simulation scenarios with variable contributions from different functional categories. We then used the six functional categories in our simulations. These categories were similar to those used in the study by Gusev et al. (2014). As in their study, we ran simulations where one of the functional categories contributed to 100% of the genetic variance, and then simulations without enrichment, where each category contributed proportionally to the number of variants present in the category. In addition, we simulated three more complex scenarios in which the different functional categories had variable contributions (Table 4.1). For these simulations, repeated 100 times per scenario, the heritability was set to 0.70 and we selected 2,000 CVs variants. In the scenarios where a single class contributed to 100%

of the heritability, the number of CVs was reduced to 500, as the number of SNPs in certain categories was limited.

4.3.5 Evaluation metrics

For each scenario, we reported summary statistics (mean, median, standard deviation, quantiles, minimum, maximum), measures of precision and accuracy (Root Mean Square Error – RMSE, and bias) of the estimators. We also reported the number of simulations without convergence with the AI-REML and after 500 additional iterations of the EM-REML.

4.3.6 Application to real data

Finally, we applied the approach to the five muscular development traits and height measured on the ~15,000 genotyped Belgian Blue beef cows and imputed to the whole-genome sequence level. We used a MC model with the same partitioning of the genome as in the simulation, except that UTR was merged with CDS (as the variability of estimates in the small category was too high). For the GREML approach, GRMs were computed using the rules described above (Yang et al., 2011a) or the first rule proposed by VanRaden et al. (2008). In addition, we also estimated the %SNP heritability associated with the different annotation classes using a TC approach.

4.4 Results

4.4.1 Estimation of proportion of genetic variance associated with a single annotation class

We first assessed whether the approaches could estimate the proportion of genetic variance associated with a specific category (also referred to as %SNP heritability) with TC models. For this purpose, we selected variants located in open chromatin regions (OCR) identified by ATAC-Seq (Yuan et al., 2023), which account for approximately 10% of the genome, and started with simulations in which these variants accounted for 50% of the genetic variance. The architecture was independent of both MAF and LD scores (i.e. CVs were randomly sampled within OCR and non-OCR). In Figure 4.1, we show the proportion of genetic variance estimated with GREML or with the BayesRR-RC model (without correction for LDMS (noLDMS), MAF-stratified (MS), LD-stratified (LDS) or LD- and MAF-stratified (LDMS) approaches). Results for each scenario are provided in Additional File 1, including summary statistics, measures of precision and accuracy, and convergence information.We observed that the %SNP heritability associated with OCR was accurately estimated with the different GREML approaches (mean = 49.7% (noLDMS), 49.7% (MS), 49.7% (LDS) and 50.4% (LDMS)), although with relatively high imprecision of the estimators (RMSE = 5.4 (noLDMS), 5.7 (MS), 5.7 (LDS) and 5.6 (LDMS)) (Additional file 1: Table S1). For example, the estimated %SNP heritability ranged from 35.9 to 67.1% when running GREML without correction for LDMS (95% of the values were between 40.4

and 58.5%). BayesRR-RC also produced estimates close to the simulated values, but with slightly higher levels of variation than GREML with the noLDMS, MS and LDS approaches. In this first scenario, the bias was below 1% with both methods, except with BayesRR-RC for the LDMS approach.

Next, we investigated whether the methods were robust to MAF- or LD-dependent architectures (MS and LDS simulations, respectively). To this end, we performed simulations in which CVs were enriched in specific MAF classes (e.g., MAF ≤ 0.05 or MAF > 0.20), LD-score categories (i.e., SNP with LD score in the lower or upper quartile), or in combinations of both features (LDMS simulation scenarios). Although the noLDMS-GREML approach provided unbiased estimates of OCR %SNP heritability in some scenarios, such as the low MAF (Figure 4.2A) and high MAF (Additional file 2: Figure S1A) scenario, high levels of bias were observed when CVs were enriched in certain LD classes (Figure 4.2B-C and Additional file 2: Figure S1B). LDS-GREML was biased in MS simulations and vice versa. Overall, only LDMS models were robust in most scenarios (Figure 4.2A-C; Additional File 1: Tables S2-7; Additional file 2: Figure S1), in agreement with previous studies (Yang et al., 2017, 2015). In this case, the estimators obtained with BayesRR-RC deviated more from the simulated values than the GREML approach. However, convergence was not systematically achieved with the GREML approach (with both the AI-REML algorithm and after 500 iterations of the EM-REML algorithm). This occurred mainly with the LDMS-GREML (Additional file 1: Tables S1-7), when a higher number of GRMs was fitted, and has also been reported in previous studies (Finucane et al., 2015; Speed et al., 2017).

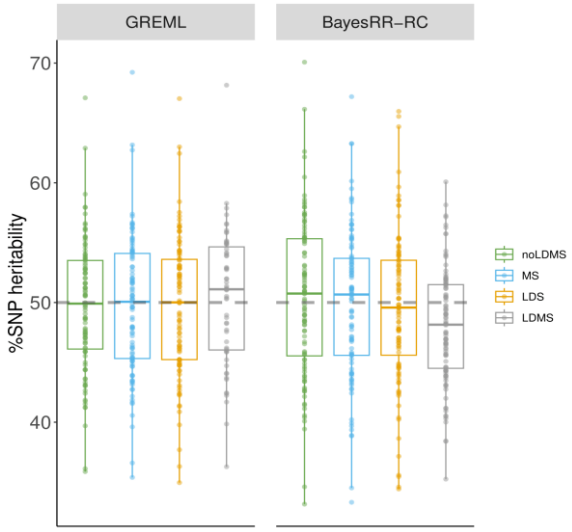


Figure 4.1. Estimation of %SNP heritability when variants in open chromatin regions (OCR) accounted for 50% of heritability. There was no additional MAF (MS) or LD stratification (LDS) in the simulations. The %SNP heritability was estimated with GREML and BayesRR-RC. The methods were applied without correction for MAF or LD score (noLDMS), and with MAF stratified (MS), LD stratified (LDMS) and both MAF and LD stratified (LDMS) approaches.

In these first simulations, each CV had the same expected contribution to the genetic variance because its effect variance was proportional to the inverse of p_j (1- p_j) (where p_j is the reference allele

frequency at SNP *j*). This architecture is consistent with the default rule used to construct the GRM in GCTA (i.e., the same architecture was used in the simulation and in the partitioning approach). We also investigated whether the accuracy of heritability partitioning would be different if different rules were used to simulate the CV effects and to construct the GRMs used in the partitioning approach. Therefore, we performed the partitioning using GRMs constructed with the first rules proposed by VanRaden (VanRaden, 2008), assuming that the CV effect variance is independent of allele frequency. In addition, we used these second rules to simulate a new scenario in which common variants contribute more to the genetic variance. In the analyses, we observed a modest bias with the noLDMS and LDS approaches when the rules used to estimate the GRM did not match those used in the simulation (Additional file 2: Figure S2). Interestingly, this bias could be reduced by using the MS and LDMS approaches.

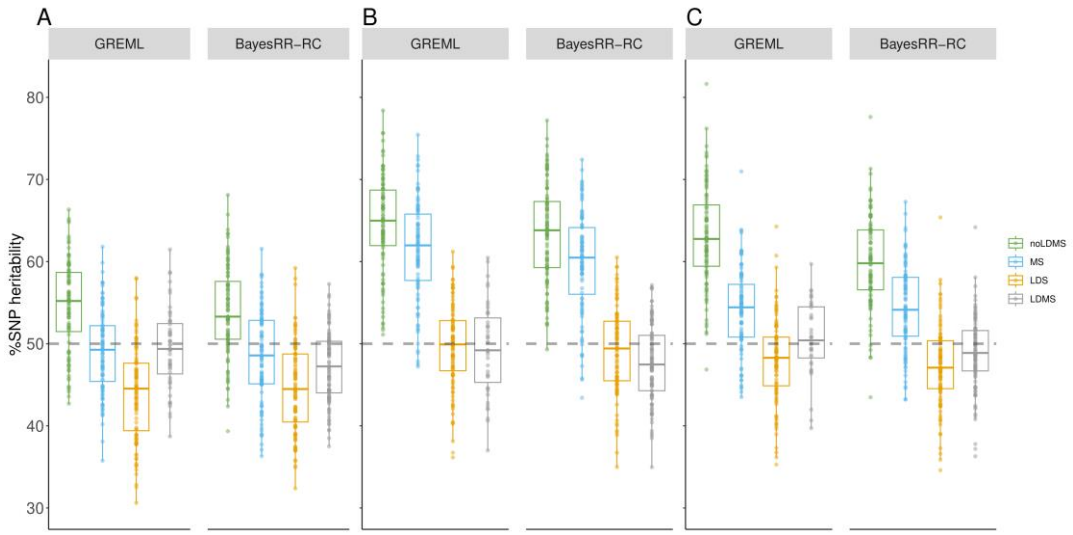


Figure 4.2. Estimation of %SNP heritability when causal variants are enriched in specific MAF or LD score categories. Variants in open chromatin regions (OCR) accounted for 50% of heritability. Causal variants were enriched in A) low MAF variants (MAF < 0.05), B) Low LD variants (LD score in the 1st quartile), and C) low MAF and low LD variants. The %SNP heritability was estimated with GREML and BayesRR-RC. The methods were applied without correction for MAF or LD score (noLDMS), and with MAF stratified (MS), LD stratified (LDMS) and both MAF and LD stratified (LDMS) approaches.

4.4.2 Estimation of proportions of the genetic variance associated with multiple annotation classes

In the second part of the study, we simulated more complex scenarios in which six different annotation classes contributed to the total genetic variance to varying degrees. The selected categories were coding regions (CDS), 3' and 5' UTR (UTRs), regions upstream and downstream of genes (+/- 1kb) called UDR, intronic regions (IOR), intergenic regions (IGR) and variants in OCR. For each simulation, we assessed whether the model was able to estimate %SNP heritability and heritability enrichment, defined as the ratio of the percentage of heritability contributed by the category to the percentage of SNPs in the category. To do this, we fitted the six categories simultaneously with a MC

model, without correcting for LDMS structure for computational reasons. We started with simulations where all the genetic variance was associated to a single class (Figure 4.3A-B; Additional file 1: Tables S8-12; Additional file 2: Figure S3). The GREML approach identified the class contributing to the genetic variation, but with relatively low precision and some bias (for instance, estimates ranged from 0.943 to 0.997 for CDS and from 0.657 to 0.979 for OCR). The BayesRR-RC approach was more accurate, with exceptionally low levels of variation in estimates across simulations, except when OCR variants accounted for 100% of the genetic variation. In this case, other categories such as CDS or UDR captured some of the variation, suggesting some confounding between these categories.

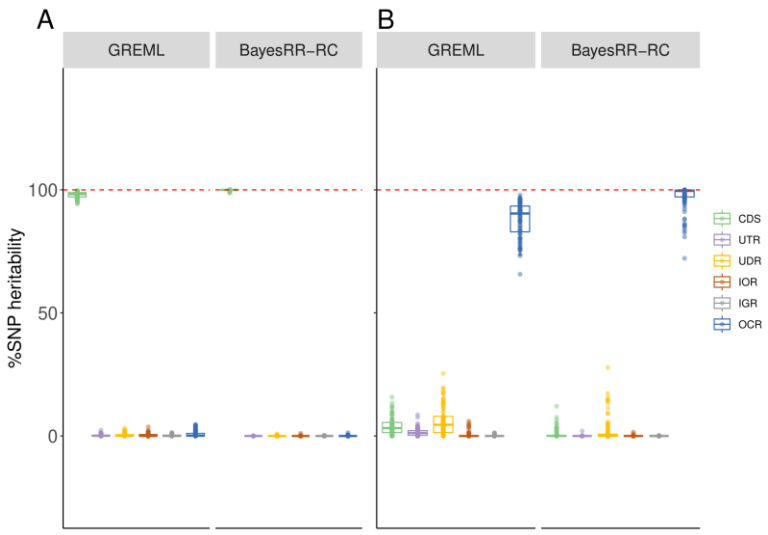


Figure 4.3. Estimation of %SNP heritability when causal variants are enriched in a single functional annotation class. Causal variants were located in A) coding sequences (CDS) and B) open chromatin regions (OCR). The %SNP heritability was estimated using GREML and BayesRR-RC with the following functional classes: CDS, 3' and 5' UTRs (UTR), upstream and downstream regions (UDR), intronic regions (IOR), intergenic regions (IGR) and OCR.

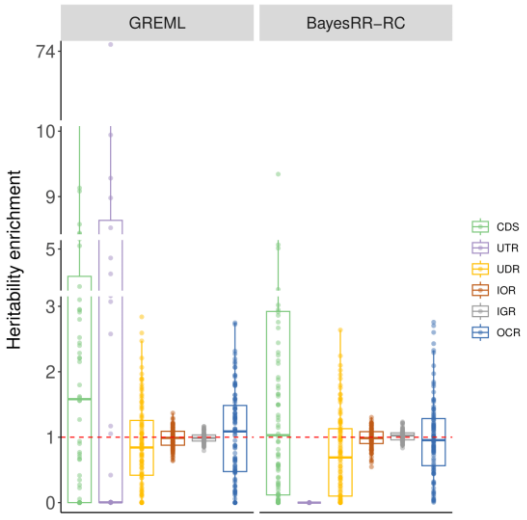


Figure 4.4. Estimation of heritability enrichment in simulations where SNPs from different functional classes had equal contribution. Heritability enrichment was estimated using GREML and BayesRR-RC with the following functional classes: coding sequence (CDS), 3' and 5' UTRs (UTR), upstream and downstream regions (UDR), intronic regions (IOR), intergenic regions (IGR) and open chromatin regions (OCR).

We then ran simulations without heritability enrichment, with the proportion of CVs per category equal to their genomic proportions. For most classes, the correct levels of enrichment were estimated by both methods (Figure 4.4; Additional file 1: Table S13), but some classes showed either high levels of variation or even some bias. The level of variation was inversely related to the size of the class, with the highest levels for the estimation of %SNP heritability for variants in UTRs and CDS. Overall, the %SNP heritability associated with each class and the ranking between classes was well estimated. We then simulated more complex and realistic scenarios with variable contributions from the different functional categories (see Table 4.1). In these scenarios, CDS and OCR were always enriched in causal variants, whereas intergenic and intronic regions harbored proportionally fewer causal variants. In the first scenario, five categories contributed 10% or more of the heritability, whereas OCR and CDS accounted for 50% or more of the genetic variation in the second and the third scenario, respectively. Results for the three scenarios are shown in Figure 4.5A-C and Additional file 1: Tables S14-16. The standard deviations of the estimators were around 0.04, but higher values were observed for OCR (over 0.08). The estimators showed some bias, with deviations generally around 0.01-0.04. The largest biases were observed for OCR and UDR, which were underestimated and overestimated respectively, confirming the confounding between these categories. In most cases, the estimators obtained with BayesRR-RC were less variable and associated with lower biases. The average RMSE, combining variation and bias, was equal to 0.063 and 0.053 for GREML and BayesRR-RC, respectively (Additional file 1: Table S17). The ranking of the different categories according to their contribution to genetic variance was not always correct, with the largest errors associated with UDR, whose contribution was systematically overestimated, and OCR. Nevertheless, the estimators provided information about which classes contributed most to genetic variation (for example, the relative importance of CDS or intergenic

variants was generally close to their simulated values). Comparisons of estimators from the same category across different scenarios (Figure 4.6) indicate that these estimators are informative despite their low precision. The coefficient of determination from the regression of estimated versus simulated values was 0.941 for CDS, 0.760 for intronic regions, 0.959 for intergenic regions and 0.804 for OCR with GREML, and 0.947 for CDS, 0.708 for intronic regions, 0.965 for intergenic regions and 0.855 for OCR with BayesRR-RC. Note that for these analyses, we did not include scenarios where classes contribute to 100% of the genetic variance, and results for UDR are not shown because its simulated values remained low in all scenarios. We repeated this analysis using estimated heritability enrichment levels (Additional file 2: Figure S4).

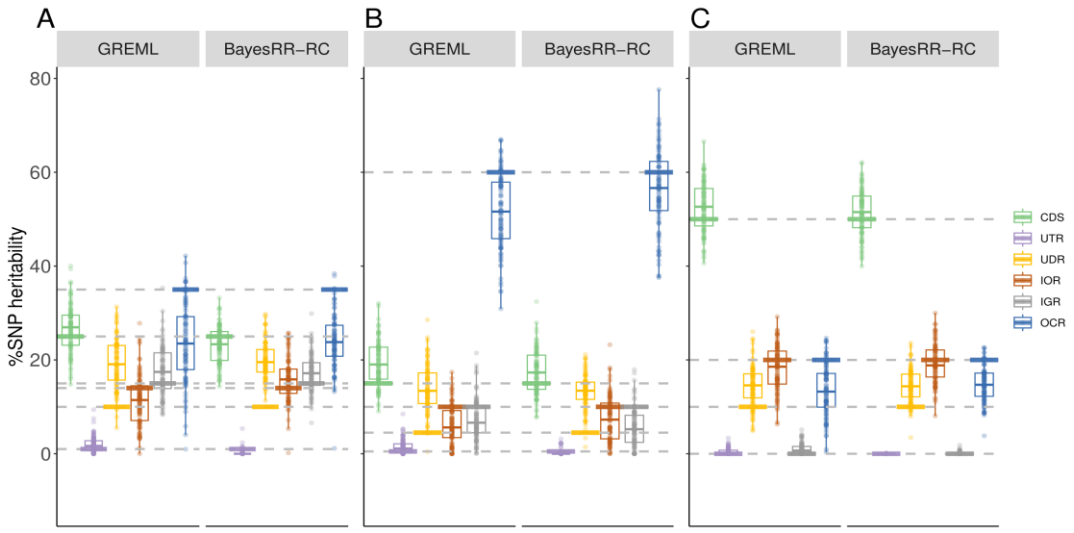


Figure 4.5. Estimation of %SNP heritability in complex simulation scenarios where SNPs from different functional classes had variable contributions. The contribution for each category is shown in Table 4.1. Heritability enrichment was estimated using GREML and BayesRR-RC with the following functional classes: coding sequence (CDS), 3' and 5' UTRs (UTR), upstream and downstream regions (UDR), intronic regions (IOR), intergenic regions (IGR) and open chromatin regions (OCR).

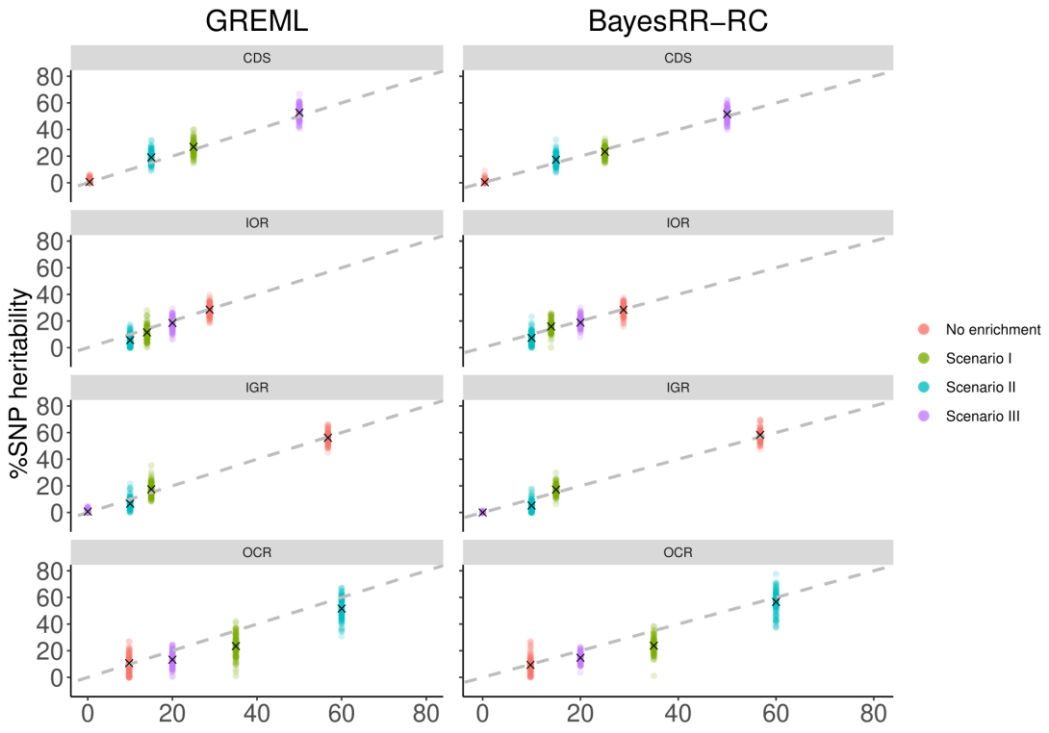


Figure 4.6. Scatterplot of estimated versus true %SNP heritability across simulation scenarios where SNPs from different functional classes contribute to genetic variance. The comparison is made separately for each functional class. %SNP heritability was estimated using GREML and BayesRR-RC with the following functional classes: coding sequence (CDS), 3' and 5' UTRs (UTR), upstream and downstream regions (UDR), intronic regions (IOR), intergenic regions (IGR) and open chromatin regions (OCR).

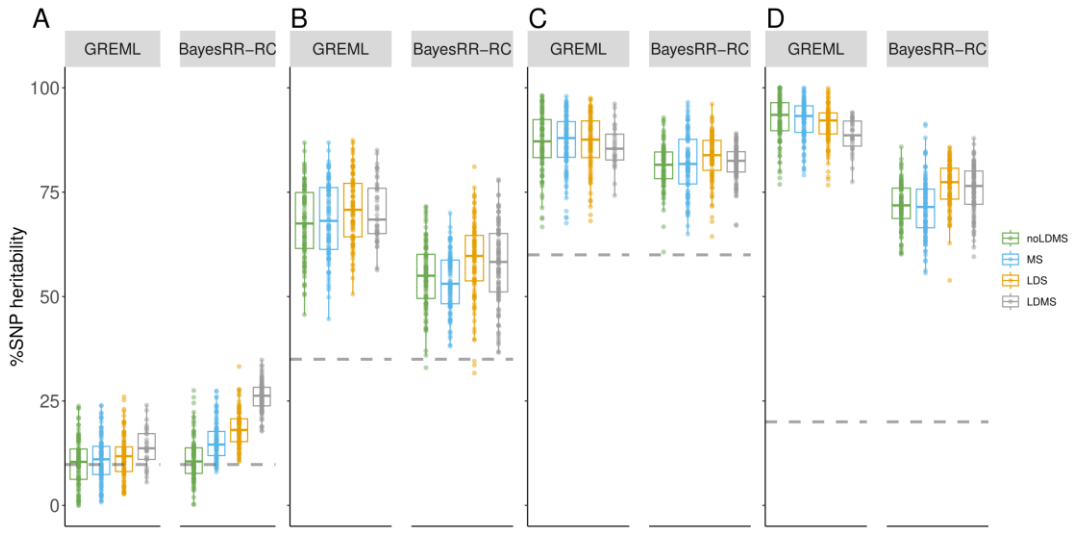


Figure 4.7. Estimation of %SNP heritability of variants in open chromatin regions (OCR) using a two-component strategy. Estimation was performed in complex simulation scenarios in which SNPs from multiple functional classes contribute to genetic variance (Panel A for the scenario without enrichment and Panels B-D for complex scenarios 1 to 3, respectively). Heritability enrichment was estimated using GREML and BayesRR-RC with the following two functional classes (OCR versus other categories). In addition, methods were run with unstratified (US), MAF stratified (MS), LD stratified (LDS) and both MAF and LD stratified (LDMS) approaches.

We then evaluated the properties of the estimators obtained with models that estimate the contribution of only one functional category, using a model that fits a second category that includes all other functional classes (TC models). This approach is commonly used because it reduces computational requirements and thus allows MS, LDS or LDMS models to be applied. The approach was evaluated in the four scenarios where several categories contribute to the genetic variance, and not for UTR as the estimator was shown to be highly inaccurate due to the small size of the category. This strategy gave poor results as %SNP heritability was most often overestimated for all categories (OCR, UDR, CDS, and IOR), even when LDMS methods were used, while biases were lower for intergenic regions (Figure 4.7A-D; Additional file 1: Table S18-21; Additional file 2: Figure S5-8). The estimators showed no bias mainly in simulations without enrichment or when the category had a null contribution in the simulation. Bias was greater for OCR than for intergenic regions. In the vast majority of cases, heritability partitioning with multiple annotation groups gave better results, for example in terms of RMSE (Additional file 1: Table S22). This can also be observed when comparing estimates for a single category across multiple scenarios (Additional file 2: Figure S9). This behavior could occur because the fitted class captured variance associated with other classes due to their similarity (for example, in terms of GRM). We measured the correlations between the off-diagonal elements from GRM of each category (Additional file 1: Table S23) and observed, for example, that the GRM from IGR variants was less correlated with other GRMs, consistent with the fact that less confounding was obtained for this category. Other GRMs were highly correlated with the exception of the UTR GRM, probably because it was the smallest category. However, the correlation between GRMs from OCR and UDR was not the

highest, even though they appeared to be the most confounded indicating that other parameters influence the confounding level. For example, relative distribution of effect sizes is probably important as we don't observe confounding when enrichment levels are uniform across categories.

4.4.3 Heritability partitioning for traits related to muscular development and height in cattle

Finally, we applied the approach to the real phenotypes, as described in Material and methods. %SNP heritabilities from the different categories were relatively variable across traits. For instance, the contributions of intergenic or CDS-UTR variants estimated by BayesRR-RC were not consistent across traits, ranging from 10 to 30% of the genetic variation (Figure 4.8; Additional file 1: Table S24). Similarly, relatively large differences were observed between BayesRR-RC and GREML estimates (for instance, the estimated %SNP heritability associated with OCR was equal to 85.5 and 45.0% for height). Nevertheless, some trends were consistent across traits and methods. OCR contributed to more than 30% of the genetic variance for all traits with BayesRR-RC (25% with GREML) and most often had the largest value of %SNP heritability (Figure 4.8). The contribution of UDR was generally low, while intergenic variants had a modest contribution despite accounting for more than 50% of SNPs and indels. As in other studies, we averaged the contributions across traits (Finucane et al., 2015; Gusev et al., 2014) (Table 2). For CDS-UTR, OCR, IGR and IOR, the average estimated contributions were similar with GREML and BayesRR-RC: over 45% for OCR, around 16-19% for CDS-UTR, 17% for IOR and 10-13% for IGR. UDR had a small contribution with both approaches, but almost zero with GREML (indicating possible problems in estimating the contribution of UDR with GREML). Except for CDS-UTR, the relative ranking of the different functional categories was consistent with both methods. In terms of heritability enrichment, some trends were also consistent (Figure 4.8; Table 2; Additional file 1: Table S25). CDS-UTR had the largest enrichment (around 25 to 30-fold), followed by OCR (around 5-fold on average), whereas intronic and intergenic variants had values below 1 (0.6-fold and 0.2-fold, respectively). Partitioning with a GREML using GRMs computed with the first rules proposed by VanRaden (VanRaden, 2008) was relatively similar to the first GREML results (Table 2; Additional file 1: Tables S24-25). The estimated contributions to heritability of CDS-UTR were on average smaller, while those of the OCR were even larger. When we repeated the heritability partitioning with TC approaches without LDMS stratification, we obtained higher contributions for all functional categories (Table 4.2; Additional file 1: Tables S24-25). For example, when using GREML, the following increases were observed: +14% for CDS-UTR, +45% for IOR, +29% for UDR, +5% for IGR and +28% for OCR. These values are 1.5 times higher or more for all categories. The sum of the contributions estimated with TC approaches corresponded to more than 200% of the total genetic variance (Table 4.2). Similar results were obtained using a TC-GREML with LDMS stratification but convergence was not systematically achieved with the GREML approach.

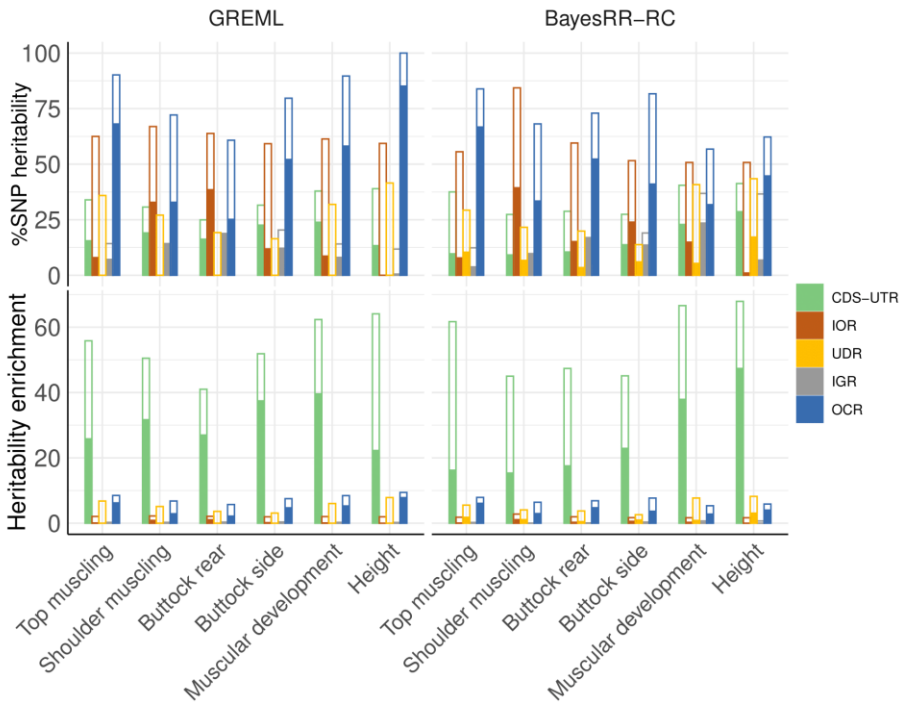


Figure 4.8. Estimation of %SNP heritability and heritability enrichment in real data sets. Estimates were obtained using GREML and BayesRR-RC with the following functional classes: coding sequence (CDS), 3' and 5' UTRs (UTR), upstream and downstream regions (UDR), intronic regions (IOR), intergenic regions (IGR) and open chromatin regions (OCR). Solid bars show %SNP heritability estimated when fitting simultaneously all the functional classes, while parameters estimated using a two-component approach, which only fits one functional category at a time, are shown with open bars.

Table 4.1. Description of the number of variants in each functional category and their contribution to SNP heritability in the three more complex scenarios. The annotations are coding sequence (CDS), intronic regions (IOR), 5' and 3' untranslated regions (UTR), up- and down-stream regions (UDR), intergenic regions (IGR) and open chromatin regions (OCR).

Annotation	Full genome		Subset of the genome		%SNP heritability		
	Number of variants	Proportion of variants	Number of variants	Proportion of variants	Scenario I	Scenario II	Scenario III
CDS	63,663	0.56%	4,161	0.43%	25	15	50
IOR	3,445,739	30.14%	278,021	28.80%	14	10	20
UTR	5,837	0.05%	412	0.04%	1	0.5	0
UDR	604,425	5.29%	40,423	4.19%	10	4.5	10
IGR	6,099,183	53.35%	547,980	56.76%	15	10	0
OCR	1,212,895	10.61%	94,429	9.78%	35	60	20

Table 4.2. Average %SNP heritability and heritability enrichment estimated for five functional groups and for six traits measured in Belgian Blue beef cattle. Values were estimated by fitting all components simultaneously with multiple classes (MC) or each component in turn with two component (TC) models and without correction for LDMS, and using BayesRR-RC or GREML (values in the parentheses correspond to the GREML partitioning when the GRMs were computed using the first rules from VanRaden (2008)).

Annotation	%SNP heritability (MC)		Heritability enrichment (MC)		%SNP heritability (TC)		Heritability enrichment (TC)	
	GREML	BayesRR-RC	GREML	BayesRR-RC	GREML	BayesRR-RC	GREML	BayesRR-RC
CDS-UTR	18.8 (14.0)	16.1	30.9 (23.0)	26.4	33.0	33.8	54.3	55.6
IOR	16.9 (16.0)	17.4	0.6 (0.5)	0.6	62.2	58.8	2.1	1.9
UDR	0.0 (1.2)	8.5	0.0 (0.2)	1.6	28.7	28.1	5.4	5.3
IGR	10.4 (9.0)	12.7	0.2 (0.2)	0.2	15.5	18.4	0.3	0.3
OCR	53.9 (59.8)	45.3	5.1 (5.6)	4.3	82.1	70.9	7.7	6.7
Total	100.0	100.0			221.5	210.0		

4.5 Discussion

4.5.1 Limitations of heritability partitioning approaches in livestock species

We herein evaluated the accuracy of GREML and BayesRR-RC in partitioning heritability according to functional classes, defined mainly on the basis of their position relative to genes and transcripts. Importantly, we evaluated the methods in a typical livestock population with reduced effective population size, high levels of relatedness and inbreeding, under intensive selection, and with high levels of long-range LD. The GREML approach has already been used in such livestock populations, for example in cattle (Bhuiyan et al., 2018; Edwards et al., 2015; Koufariotis et al., 2014; Lingzhao et al., 2017; Xiang et al., 2019a). Most often, this partitioning method was applied without an evaluation of its bias and accuracy in such context. However, differences in population structure and their impact on genome structure (e.g. LD patterns) could affect the precision and accuracy of the methods. For example, in humans, the methods have been evaluated by carefully filtering out pairs of individuals with levels of relatedness greater than 0.025 (Yang et al., 2015). In livestock, a large fraction of pairs of individuals would have levels above such a threshold. Recently, Cai et al. (2022) conducted a study to evaluate different GREML approaches for estimating heritability enrichment in a cattle population. They used data from 2,000 Holstein bulls imputed for about 700,000 markers, and mainly evaluated the accuracy of the estimators for three different MAF categories. Although some models gave unbiased results, biased estimators were observed when parameters from the simulated and fitted models did not match (Cai et al., 2022). In particular, they found that estimated enrichment values were biased when CVs were enriched in rare alleles and that using LD scores calculated in too large windows resulted in biased estimates. We herein performed a simulation approach based on a large cohort of individuals. Importantly, our data were imputed at the whole-genome sequence level, providing a finer resolution for annotation. Compared to the study by Cai et al. (2022), we included more functional annotation groups, including information from a recently published ATAC-Seq peak catalogue (Yuan et al., 2023), and we explored more scenarios (CVs could be enriched as a function of MAF, LD score and functional annotation). Using this approach, we first observed that in relatively simple scenarios (no stratification by MAF or LD, with CVs enriched for a single functional category), the methods were unbiased, but that the estimates showed high levels of variation. Note that when simulations were performed using the whole genome, even higher levels of variation were observed with the GREML approach (data not shown). When CVs were enriched in a particular MAF or LD score category, it was necessary to stratify the GREML or BayesRR-RC accordingly to obtain unbiased results (i.e., using a LDMS approach), consistent with findings in humans (Yang et al., 2015). When GRMs were not defined for different MAF or LD groups, biased partitioning was indeed obtained. Importantly, the LD or MS groups fitted in the partitioning methods should match those that are truly enriched in CVs, an information that is rarely known. Other elements could further bias the results, such as the relationship between the MAF or LD scores of CVs and the magnitude of their effects, as previously highlighted by

Speed et al. (2017) or Cai et al. (2022). For example, the fitted GRMs could assume equal SNP contribution to the genetic variance (rare alleles having then larger effects) or comparable effect sizes for all SNPs regardless of MAF (common SNPs having higher contribution to the genetic variance), whereas the true relationship between CVs and MAF could be different. Simulation results indicated that such different between simulated and fitted architecture can sometimes be compensated by the use of an LDMS approach. Next, we ran simulations in which multiple functional categories contributed to the phenotypic variation with different levels of enrichment. The estimators still showed high levels of variation, especially for the classes with few variants, but we also observed systematic biases due to confounding between some functional categories, the strongest between OCR and UDR. Estimators were better for variants in IGR, as their GRM was less similar to the GRM of other categories. This is important because it implies that confounding is higher for functional categories that are expected to contribute most to the genetic variance, and thus their estimates are less precise. We also tested a strategy estimating the %SNP heritability of each category individually (running one TC-GREML per category) and observed very strong biases, probably due to confounding. The estimated %SNP heritabilities were greatly overestimated for most categories. Although this two-component strategy reduces computational costs and allows fitting a LDMS model, it is therefore not recommended. This is an important observation as this is a common strategy (Cai et al., 2022; de las Heras-Saldana et al., 2020; Edwards et al., 2015; Koufariotis et al., 2014; Lingzhao et al., 2017; Xiang et al., 2023, 2019a).

The high levels of variation in heritability enrichment estimates could also be due to the similarity between GRMs from different functional or LDMS categories, or due to LD between neighboring SNPs from different categories. This problem is likely to be more severe in livestock species because the additive genetic relationships r_{xy} between pairs of individuals *x* and *y* are spread over a wider range, including unrelated individuals (r_{xy} = 0), half-sibs (r_{xy} = 0.25), full-sibs (r_{xy} = 0.5), parent-offspring (r_{xy} = 0.5) and even monozygotic twin (r_{xy} = 1) pairs. The high levels of relatedness will drive the correlations between elements from the different GRMs and may mask more subtle correlations due to short-range LD between SNPs. It has been shown that the properties of heritability estimators are different when individuals are unrelated and LD is high only at short distances (Campos et al., 2015). When GRMs from different fitted categories are more distant, the problem of bias due to confounding between categories is likely to be less. This would be the case, for example, in studies evaluating the contribution from each autosome separately (Bhuiyan et al., 2018; Robinson et al., 2013; Yang et al., 2011b), or from specific chromosomes of interest such as the sex chromosomes (Kadri et al., 2022). For example, GRM from sex chromosomes are based on different segregation rules and are less correlated with GRMs obtained from autosomes (Druet and Legarra, 2020; Yang et al., 2011b). Similarly, relationship matrices could be estimated for mitochondria or chloroplasts in plants to assess their contribution to the genetic variance.

4.5.2 Comparison of evaluated methods.

Our evaluation focused on two methods, GREML and BayesRR-RC. For GREML, we estimated the GRM using the rules from Yang et al. (2011a). Different GRM construction rules can lead to different estimators. For example, the GRM may be based on different relationships between the variance of marker effects and their MAF (by defining a parameter called "α"), LD score, or their genotyping accuracy (Speed et al., 2017, 2012). The accuracy of GREML with different values of α has previously been evaluated in a livestock population by Cai et al. (2022). In preliminary tests, we obtained less accurate estimates with the LDAK-Thin model recommended for non-human organisms (Speed et al., 2012), and therefore selected original rules from Yang et al. (2011a), assuming that each marker has an equal expected contribution to heritability (i.e., independent of MAF), to construct our GRM. Nevertheless, in additional simulations, we observed that a mismatch between the function used to compute the GRM and the simulated relationship between CV effects and their MAF or LD scores could bias the results, more so when the relationship with LD scores was suboptimal and with GREML (data not shown). Unfortunately, the relationships remain unknown and, based on our results, the use of BayesRR-RC and LDMS models is recommended in such situations. The LD regression score (LDRS) is another method that allows heritability partitioning (Finucane et al., 2015). It is computationally efficient because it relies on summary statistics. Nevertheless, heritability estimates from LDRS have higher standard errors than those from GREML (Bulik-Sullivan, 2015; Speed et al., 2017), and the approach has not been shown to be more efficient than GREML or BayesRR-RC in several studies (Patxot et al., 2021; Speed and Balding, 2019). The properties of LDRS need to be evaluated in livestock populations, where the extent of LD is very different from humans and where markers may be in linkage equilibrium with CVs due to the presence of high levels of relatedness. For instance, Xiang et al. (2023) obtained poor results with LDRS in dairy cattle. In addition, obtaining summary statistics in livestock populations is more computationally demanding because LMM must be used for GWAS to correct for stratification and polygenic background. Due to these high computational requirements and based on previous comparison results, we did not evaluate LDRS in our study. Compared to BayesRR-RC, GREML produced more accurate results in the first set of simulations where OCR variants accounted for 50% of the heritability. In similar cases, LDMS models are recommended to obtain unbiased results. However, with many different fitted components, 500 iterations of the EM algorithm were sometimes insufficient to achieve convergence. These problems could be reduced by fitting a two-component model, but this produced biased results (see above). When we fitted models with multiple functional categories, BayesRR-RC outperformed the GREML approach. However, Bayes-RR-RC has higher computational costs and the number of iterations that can be run is relatively small. Convergence diagnostic plots and comparisons with longer chains suggest that this number of iterations already provides good estimates for most parameters although these had high levels of variation (see Additional file 2: Figures S10 and S11). This is consistent with the results of Orliac et al. (2022) who concluded that less than 5,000 iterations are required to estimate variance components and for genomic predictions.

In the most complex scenarios, the estimator for some parameters was not fully stabilized after 5,000 iterations (Additional file 2: Figure S11). This suggests that more iterations may be required for livestock species due to the higher LD and relatedness levels. Nevertheless, comparisons of the results obtained with 5,000 versus 50,000 iterations for 25 simulations from 2 scenarios show that the distributions of the estimated parameters are very similar. Overall, we observed that, with a total of 5,000 iterations, Bayes-RR-RC performed better than GREML, but we cannot exclude that longer chains could further improve the results.

4.5.3 Heritability partitioning for muscularity and height in Belgian Blue beef cattle

Despite the high standard errors in the simulations, the estimated heritability enrichments and their ranking remain informative, especially when averaged over multiple traits, as done in other studies (Finucane et al., 2015; Gusev et al., 2014). With both GREML and BayesRR-RC, variants present in OCR had by far the largest contribution to heritability (> 45%). Regulatory regions have also been shown to have the largest contribution to genetic variance for complex traits in humans (Gusev et al., 2014) and to be important in cattle (Koufariotis et al., 2014; Xiang et al., 2019a). Recently, Xiang et al. (2023) evaluated that regulatory variants explained up to 70% of the genetic variance in cattle. In terms of heritability enrichment, coding variants had the highest average per-variant contribution to the heritability (> 25-fold on average), variants in the OCR also showed substantial enrichment (~5-fold), whereas intronic and intergenic variants had enrichment values below 1 (0.6 and 0.2-fold, respectively). This ranking is in line with expectations and is consistent with results obtained in several studies of complex traits in humans (Finucane et al., 2015; Gusev et al., 2014). The observation of large effects associated with coding variants is in agreement with the findings of Gualdron Duarte et al. (2023), who identified several such variants associated with the same traits and accounting for a large proportion of the genetic variance. Heritability partitioning could be refined by using more specific functional classes such as coding variants or eQTLs, but care must be taken as we have shown the limitations of partitioning approaches when too small or too many categories were fitted. Similarly, heritability enrichment could be applied to other types of categories such as conservation scores, differentiation scores, evidence of selection, or age of alleles.

4.6 Conclusions

Here we have shown that heritability partitioning approaches should be used cautiously in livestock populations and that accuracy assessment is strongly recommended. Estimators were particularly imprecise for small categories, so models with too many and small functional categories should not be used. In addition, two-component approaches that fit only one functional category at a time produced biased estimates and should not be used. Nevertheless, the estimates and their ranking were still informative about the contribution of the functional classes we fitted. We therefore applied

the methods to real phenotypes for muscular development and height. We estimated that, on average, variants in open chromatin regions had a higher contribution to the genetic variance, while variants in coding regions had the strongest individual effects. Conversely, variants in intergenic or intronic regions showed lower levels of enrichment. The results are consistent with those obtained in humans.

4.7 Acknowledgements

The authors acknowledge the Walloon Breeders Association (AWE group) for providing the data. Tom Druet is Research Director from the F.R.S.-FNRS. We used the supercomputing facilities of the "Consortium d'Equipements en Calcul Intensif en Fédération Wallonie-Bruxelles" (CECI), funded by the F.R.S.-FNRS.

4.8 Supplementary tables

https://figshare.com/articles/dataset/Additional_file_1_of_Evaluation_of_heritability_partitioning_approaches_in_livestock_populations/26299302?file=47669655

4.9 Supplementary figures

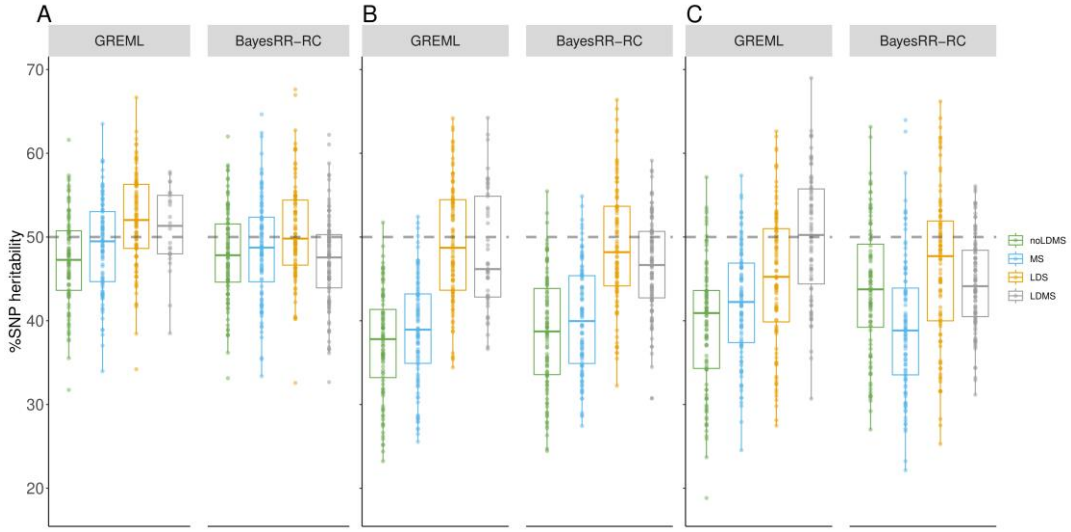


Figure S1. Estimation of %SNP heritability when causal variants are enriched in specific MAF or LD score categories. Variants in open chromatin regions (OCR) accounted for 50% of heritability. Causal variants were enriched in A) common variants (MAF > 0.20), B) high LD variants (LD score above the 3rd quartile), and C) low MAF (MAF <0.05) and high LD (LD score above the 3rd quartile) variants. The %SNP heritability was estimated with GREML for simulations using the whole genome (GREML – FULL) and with GREML and BayesRR-RC for simulations using a subset of the genome. The methods were applied without correction for MAF or LD score (noLDMS), and with MAF stratified (MS), LD stratified (LDS) and both MAF and LD stratified (LDMS) approaches.

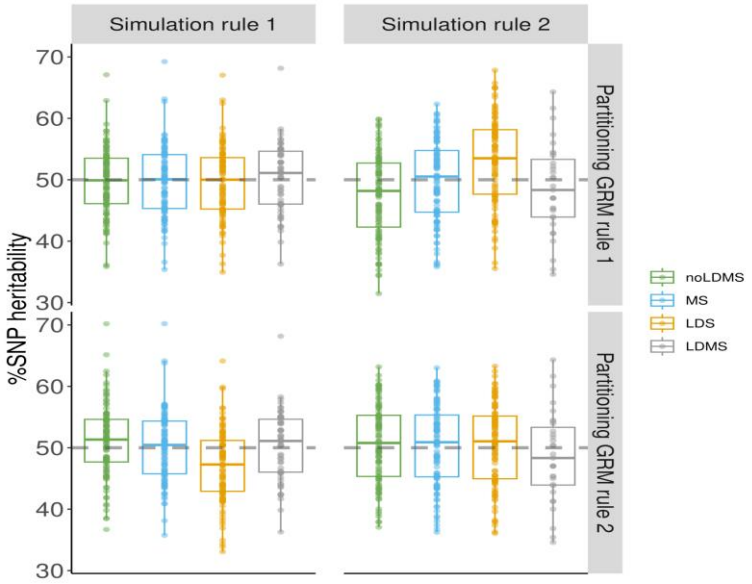


Figure S2. Estimation of %SNP heritability using different GRM computation methods and for the two scenarios where SNP effect size is a function of allele frequency. Simulation rule 1: SNP effects increase as allele frequencies decrease (corresponding to the default rule). Simulation rule 2: SNP effects are drawn from the same distribution regardless of allele frequency (corresponding to the rules proposed by VanRaden [29]). Partitioning GRM rule 1: GRMs used in the heritability partitioning are computed using the default rules from GCTA. Partitioning GRM rule 2: GRMs used in heritability partitioning are computed using the VanRaden rules from.

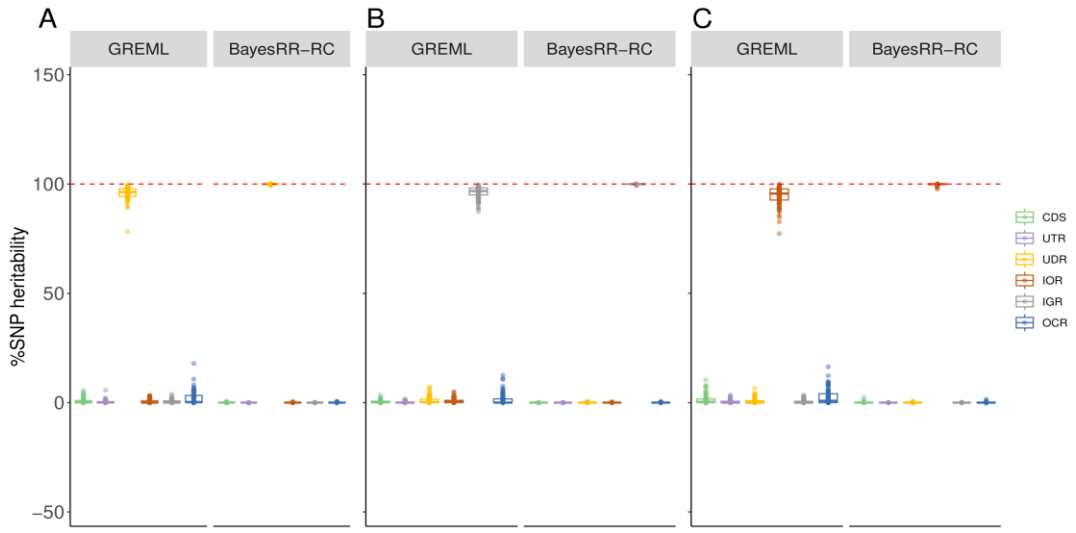


Figure S3. Estimation of %SNP heritability when causal variants are enriched in a single functional annotation class. Causal variants were located in A) upstream and downstream regions (UDR), B) intergenic regions (IGR), and C) intronic regions (IOR). The %SNP heritability was estimated using GREML and BayesRR-RC with the following functional classes: coding sequence (CDS), 3' and 5' UTRs (UTR), UDR, IOR, IGR and open chromatin regions (OCR).

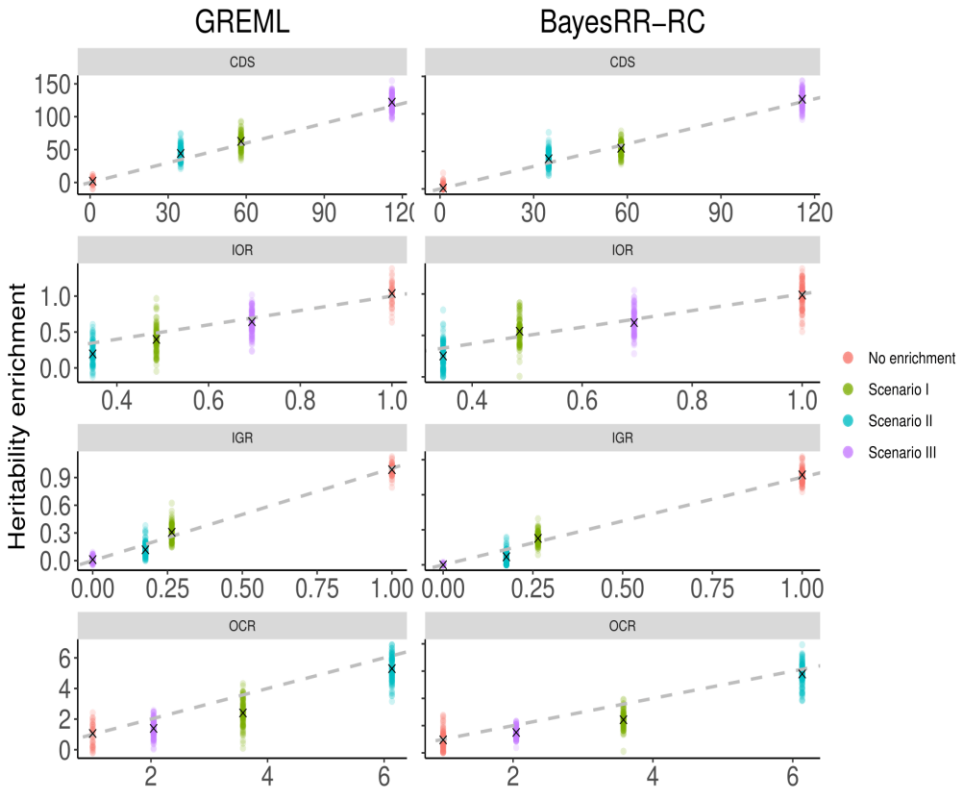


Figure S4. Scatterplot of estimated versus true heritability enrichment across simulation scenarios where SNPs from different functional classes contribute to genetic variance. The comparison is made separately for each functional class. Heritability enrichment was estimated using GREML and BayesRR-RC with the following functional classes: coding sequence (CDS), 3' and 5' UTRs (UTR), upstream and downstream regions (UDR), intronic regions (IOR), intergenic regions (IGR) and open chromatin regions (OCR).

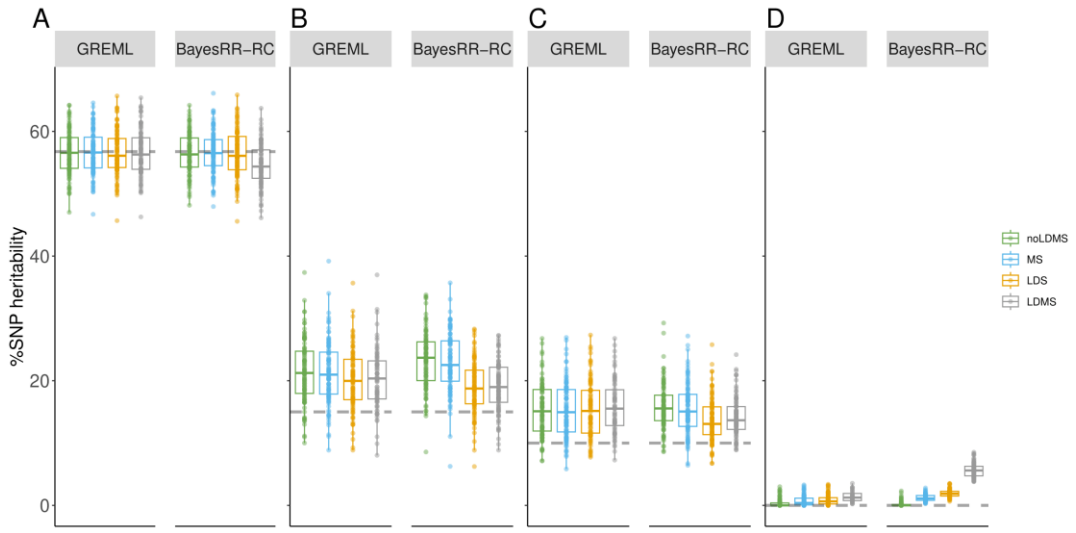


Figure S5. Estimation of %SNP heritability of variants in intergenic regions (IGR) using a two-component strategy. Estimation was performed in complex simulation scenarios in which SNPs from multiple functional classes contribute to genetic variance (Panel A for the scenario without enrichment and Panels B-D for complex scenarios 1 to 3, respectively). Heritability enrichment was estimated using GREML and BayesRR-RC with the following two functional classes (IGR versus other categories). In addition, methods were run with unstratified (US), MAF stratified (MS), LD stratified (LDS) and both MAF and LD stratified (LDMS) approaches.

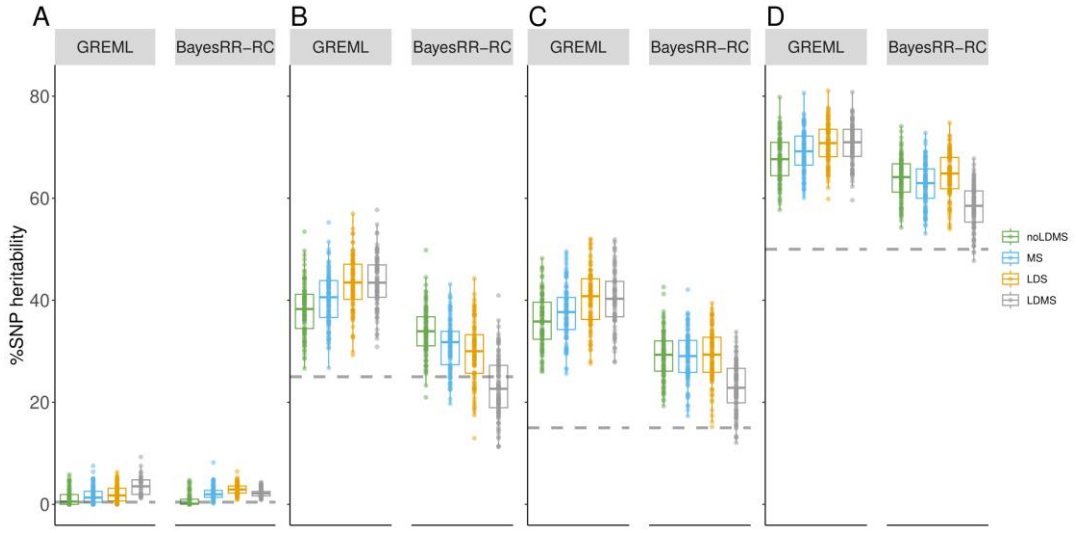


Figure S6. Estimation of %SNP heritability of variants in coding sequence (CDS) using a two-component strategy. Estimation was performed in complex simulation scenarios in which SNPs from multiple functional classes contribute to genetic variance (Panel A for the scenario without enrichment and Panels B-D for complex scenarios 1 to 3, respectively). Heritability enrichment was estimated using GREML and BayesRR-RC with the following two functional classes (CDS versus other categories). In addition, methods were run without correction for MAF or LD score (noLDMS), and with MAF stratified (MS), LD stratified (LDS) and both MAF and LD stratified (LDMS) approaches.

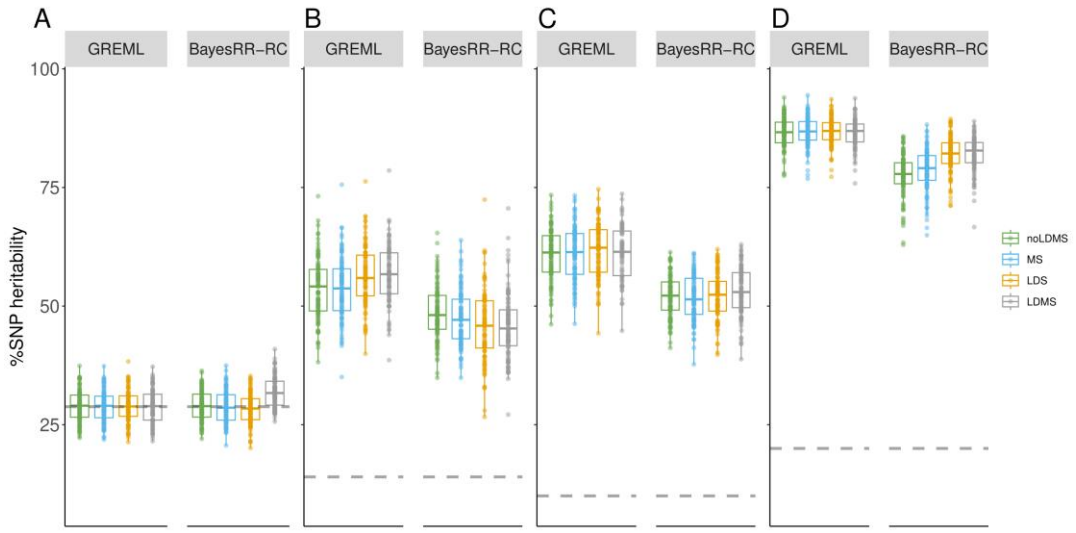


Figure S7. Estimation of %SNP heritability of variants in intronic regions (IOR) using a two-component strategy. Estimation was performed in complex simulation scenarios in which SNPs from multiple functional classes contribute to genetic variance (Panel A for the scenario without enrichment and Panels B-D for complex scenarios 1 to 3, respectively). Heritability enrichment was estimated using GREML and BayesRR-RC with the following two functional classes (IOR versus other categories). In addition, methods were run without correction for MAF or LD score (noLDMS), and with MAF stratified (MS), LD stratified (LDS) and both MAF and LD stratified (LDMS) approaches.

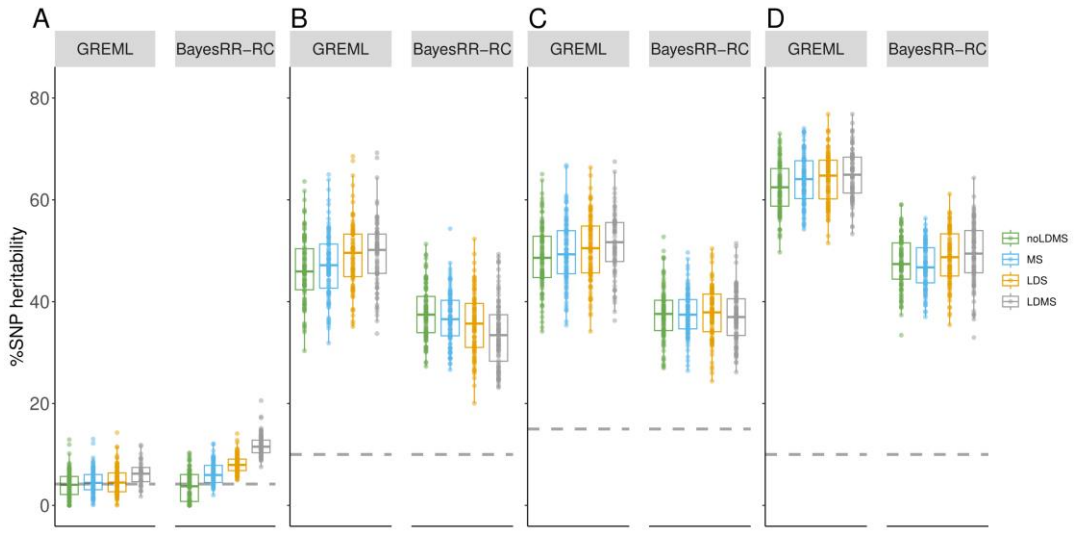


Figure S8. Estimation of %SNP heritability of variants in upstream and downstream regions (UDR) using a two-component strategy. Estimation was performed in complex simulation scenarios in which SNPs from multiple functional classes contribute to genetic variance (Panel A for the scenario without enrichment and Panels B-D for complex scenarios 1 to 3, respectively). Heritability enrichment was estimated using GREML and BayesRR-RC with the following two functional classes (UDR versus other categories). In addition, methods were run without correction for MAF or LD score (noLDMS), and with MAF stratified (MS), LD stratified (LDS) and both MAF and LD stratified (LDMS) approaches.

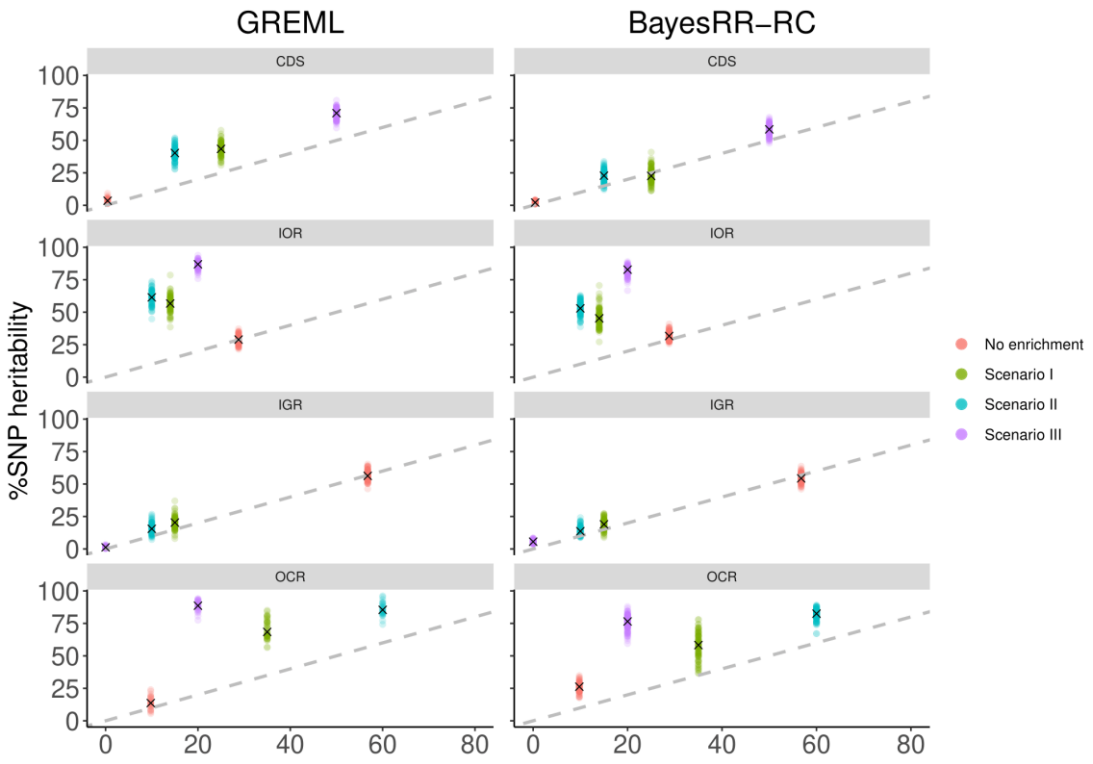


Figure S9. Scatterplot of estimated versus true %SNP heritability when using a two-component strategy. Estimates were compared across simulation scenarios where SNPs from different functional classes contribute to genetic variance. The contribution for each category is shown in Table 4.1. The comparison is made separately for each functional class. %SNP heritability was estimated using GREML and BayesRR-RC with the following two functional classes (one versus other categories) and a MAF and LD stratified (LDMS) approach. Fitted functional categories were coding sequence (CDS), 3' and 5' UTRs (UTR), upstream and downstream regions (UDR), intronic regions (IOR), intergenic regions (IGR) and open chromatin regions (OCR).

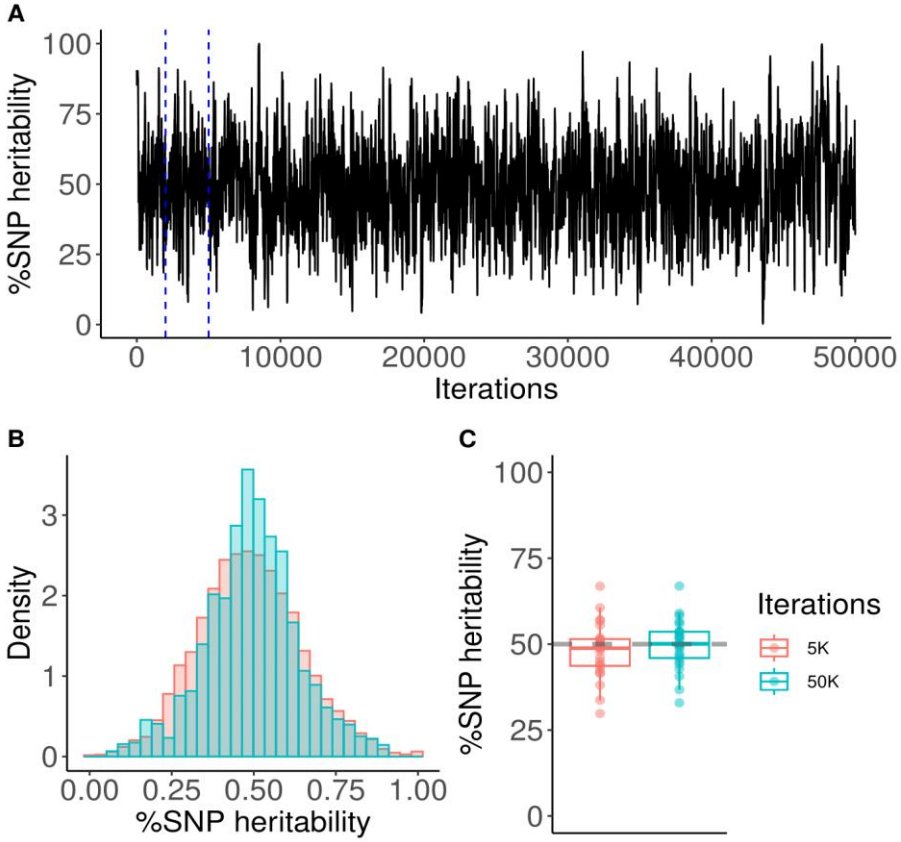


Figure S10. Comparison of BayesRR-RC results obtained with 5,000 versus 50,000 iterations in a simple scenario. The model was run on data from a simple scenario where OCR contributed to 50% of the genetic variance. The 5,000 iterations correspond to the values used in the present study (burn-in from iterations 1-2,000), while 50,000 iterations correspond to a longer run (burn-in from iterations 1-5,000). A) Estimated %SNP heritability per iteration. Iterations used for parameter estimation in the standard run are delimited by the two blue dashed lines located at iterations 2,001 and 5,000. B) Distribution of %SNP heritability estimates in iterations 2,001-5,000 (standard run) and 5,001-50,000 (long run). C) %SNP heritability estimates for 25 simulations estimated using BayesRR-RC with 5,000 versus 50,000 iterations.

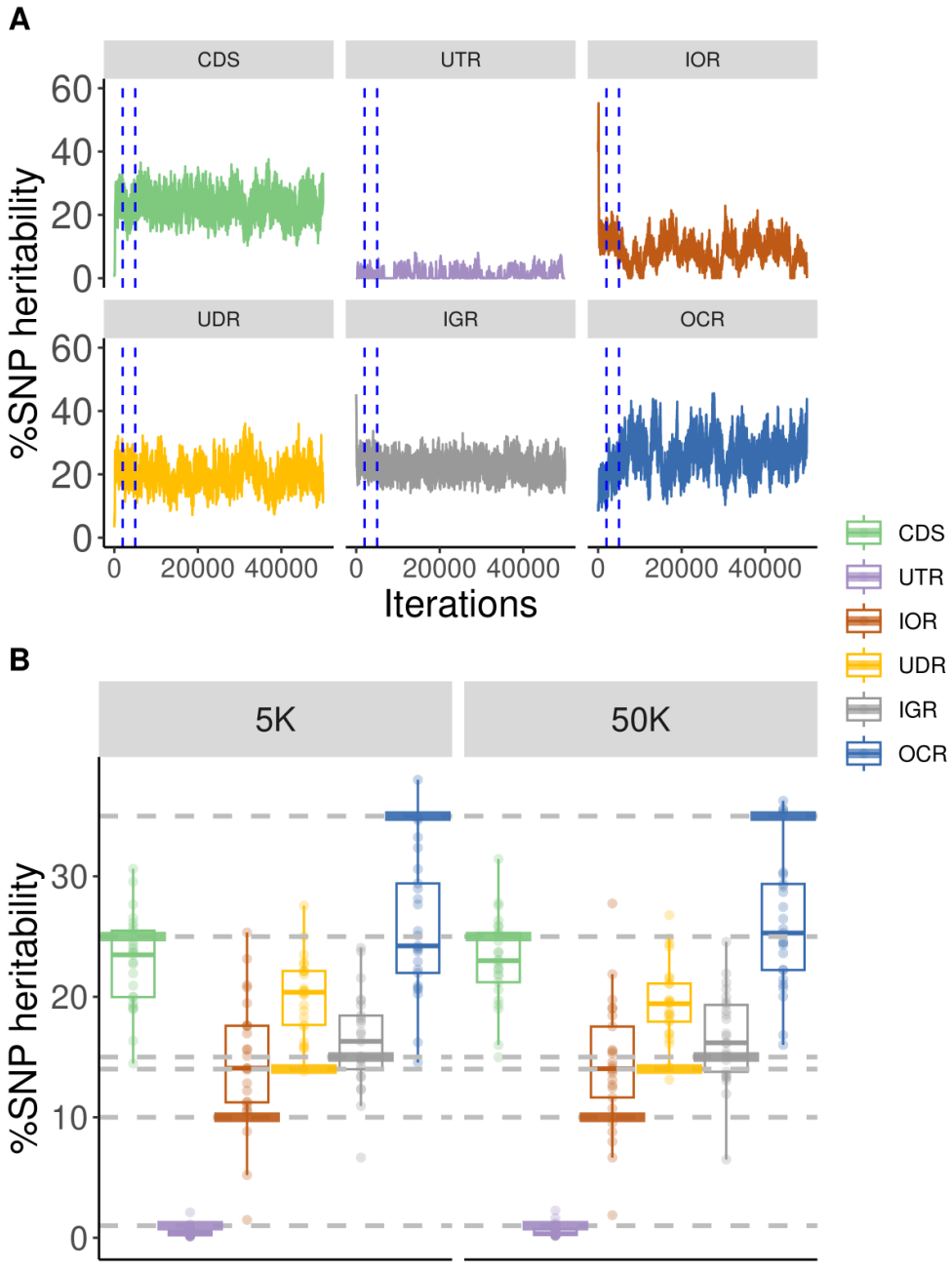


Figure S11. Comparison of BayesRR-RC results obtained with 5,000 versus 50,000 iterations in the first complex scenario. The 5,000 iterations correspond to the values used in the present study (burn-in from iterations 1-2,000), while 50,000 iterations correspond to a longer run (burn-in from iterations 1-5,000). A) Estimated %SNP heritability per iteration for the six components. Iterations used for parameter estimation in the standard run are delimited by the two blue dashed lines located at iterations 2,001 and 5,000. B) %SNP heritability estimates for the six components estimated using BayesRR-RC with 5,000 versus 50,000 iterations in 25 simulations.

Experimental section

Study 3

Evaluation of genomic selection models using whole genome sequence data and functional annotation in Belgian Blue cattle

Genetics Selection Evolution - Submitted

Can Yuan, Alain Gillon, José Luis Gualdrón Duarte, Haruko Takeda, Wouter Coppieters, Michel Georges and Tom Druet

5 Experimental section: Study 3

5.1 Summary

The availability of large cohorts of whole-genome sequenced individuals, combined with functional annotation, is expected to provide opportunities to improve the accuracy of genomic selection (GS). However, such benefits have not often been observed in initial applications. The reference population for GS in Belgian Blue Cattle (BBC) continues to grow. Combined with the availability of reference panels of sequenced individuals, it provides an opportunity to evaluate GS models using whole genome sequence (WGS) data and functional annotation. Here, we used data from 16,508 cows, with phenotypes for five muscular development traits and imputed at the WGS level, in combination with in silico functional annotation and catalogs of putative regulatory variants obtained from experimental data. We evaluated first GS models using the entire WGS data, with or without functional annotation. At this marker density, we were able to run two approaches, assuming either a highly polygenic architecture (GBLUP) or allowing some variants to have larger effects (BayesRR-RC, a Bayesian mixture model), and observed an increased reliability compared to the official GBLUP model at medium marker density (on average 0.016 and 0.018 for GBLUP and BayesRR-RC, respectively). When functional annotation was used, we observed slightly higher reliabilities with an extension of the GBLUP that included multiple polygenic terms (one per functional group), while reliabilities decreased with BayesRR-RC. We then used large subsets of variants selected based on functional information or with a linkage disequilibrium (LD) pruning approach, which allowed us to evaluate two additional approaches, BayesCπ and Bayesian Sparse Linear Mixed Model (BSLMM). Reliabilities were higher for these panels than for the WGS data, with the highest accuracies obtained when markers were selected based on functional information. In our setting, BSLMM systematically achieved higher reliabilities than other methods. GS with large panels of functional variants selected from WGS data allowed a significant increase in reliability compared to the official genomic evaluation approach. However, the benefits of using WGS and functional data remained modest, indicating that there is still room for improvement, for example by further refining the functional annotation in the BBC breed.

5.2 Introduction

The implementation of genomic selection (Meuwissen et al., 2001) in livestock species has been made possible by the development of high-throughput genotyping technologies. Indeed, the availability of low-cost genotyping arrays has led to the rapid adoption of genomic selection in many livestock species and breeds (Meuwissen et al., 2016). However, the availability of whole genome sequence (WGS) should make it possible to further improve the accuracy of genomic selection, as causative variants would be included in the model. Furthermore, with sequence-based genomic selection, the accuracy of predictions would remain high over multiple generations, as the linkage disequilibrium (LD) between markers and causative variants would not decay over generations. With the ability to sequence large reference panels of individuals (Daetwyler et al., 2014; Ros-Freixedes et al., 2022) and the availability of efficient genotype phasing and imputation tools (Browning et al., 2018; Das et al., 2016; Delaneau et al., 2019; Rubinacci et al., 2020), it is becoming increasingly common to have imputed WGS data for large cohorts of individuals. However, in early studies based on either simulated or real data, the use of imputed WGS data resulted in no or small improvements in prediction accuracy when the prediction methods were not changed (Druet et al., 2014b; Frischknecht et al., 2018; Pérez-Enciso et al., 2015; van Binsbergen et al., 2015; Veerkamp et al., 2016), whereas predictions using only the causative variants provided a significant improvement (Pérez-Enciso et al., 2015), especially when they were rare (Druet et al., 2014b).

To fully exploit the potential of whole genome sequence information, other strategies are needed. Two main directions have been proposed in the literature: 1) using additional information to classify variants into different functional categories having different effect sizes; 2) selecting a subset of markers that are more likely to be causative from the whole-genome sequence data, either to reduce model dimensionality or to add the markers to custom genotyping arrays. Two main groups of methods developed to apply the first strategy are commonly used. The first group includes extensions of the genomic best linear unbiased prediction (GBLUP) that fit multiple polygenic terms with their own genomic relationship matrix (GRM), such as the genomic feature BLUP (GFBLUP) (Edwards et al., 2016; Sørensen et al., 2014). With the GFBLUP, annotation groups are fitted one by one (next to a polygenic term that fits the rest of the genome), but models that fit more than two annotation groups are possible, as in the MultiBLUP model (Speed and Balding, 2014). The second group includes extensions of the BayesR model (Erbe et al., 2012), a Bayesian mixture of Gaussian distributions associated with different SNP effect sizes, including BayesRC (MacLeod et al., 2016), BayesRCO (Mollandin et al., 2022) and BayesRR-RC (Patxot et al., 2021) models. With both approaches, genetic variants are classified into different annotation groups, which may have group-specific parameters such as effect variances or mixture parameters. The GFBLUP approach has been used to perform heritability partitioning and genomic prediction using different features, such as genome-wide association studies (GWAS) results, expression QTLs (eQTL) (Ehsani et al., 2016) and Gene Ontology categories

(Lingzhao et al., 2017). GFBLUP was found to be more accurate than GBLUP in several studies, although not systematically. McLeod et al. (2016) used BayesRC using annotation categories related to coding and putative regulatory variants, specific to lactation genes or not, and achieved slightly higher accuracies compared to a traditional BayesR model. A method similar to BayesRC has also been shown to be efficient for predicting complex traits in humans (Orliac et al., 2022; Patxot et al., 2021). However, these annotation-aware approaches have rarely been applied to complete whole-genome sequence data in livestock species (especially the Bayesian approach), and strategies based on marker pre-selection are often implemented. After this marker selection step, genomic predictions can be applied with or without grouping variables according to relevant features. It is common to select markers based on GWAS results (Brøndum et al., 2015; Frischknecht et al., 2018; A. Liu et al., 2020a; Ros-Freixedes et al., 2022; VanRaden et al., 2017; Veerkamp et al., 2016), but other criteria have also been used such as coding variants (Frischknecht et al., 2018; A. Liu et al., 2020a), eQTL (de las Heras-Saldana et al., 2020), putative regulatory regions (A. Liu et al., 2020a) or more general genomic annotations based on position relative to genes (VanRaden et al., 2017). Xiang et al. (2019a) used probably the most complete set of criteria in cattle, including functional and evolutionary information, and proposed a global score for each marker. Finally, although most of the time genotypes from selected markers are imputed, there are sometimes included on custom genotyping arrays (Khansefid et al., 2020; A. Liu et al., 2020a).

The main objective of the present study was to evaluate strategies to improve the accuracy of genomic selection in Belgian Blue cattle (BBC) using imputed whole genome sequence data and functional annotation. This breed is mainly selected for muscular development traits, with the fixation of an 11-bp deletion in the myostatin (*MSTN*) gene associated with double muscling. Recent studies have improved our knowledge of the genetic architecture of these traits. First, selective sweeps revealed that large effect variants have been fixed by selection (Druet et al., 2014a, 2013), but only two of the identified hard sweeps were associated with complex traits, and only one was breed-specific and related to muscularity (the *MSTN* mutation). This is consistent with the review by Kemper and Goddard (2012), who stated that most loci associated with complex traits in cattle have small effects, but that variants with larger effects can occasionally segregate in the population. Next, a recent sequence-based GWAS study (Gualdrón Duarte et al., 2023) showed that the significant associations are enriched for common coding variants with large effects. However, these correspond to a relatively small number of variants (< 15), those with the largest effects, and contribute only to a small proportion of the genetic variance. In line with this, Yuan et al (2024) estimated that putative regulatory variants have the highest contribution to heritability and that coding variants have the highest enrichment levels (i.e. have the largest effects on average). The high contribution of regulatory variants is consistent with the findings of Xiang et al. (2023), who recently estimated that gene expression and RNA splicing explain large proportions of the heritability for complex traits. Therefore, we will place more emphasis on coding and regulatory variants in the strategies evaluated. A particular focus will be on putative regulatory elements detected in muscle, as the breed is primarily selected for muscular development.

5.3 Material and methods

5.3.1 Data

Our study used a cohort of 18,324 BBC genotyped cows that we imputed at the sequence level. The genotyping data and methodology are very similar to those described in Gualdrón Duarte et al. (2023), where further details can be found. Briefly, cows were genotyped on 10 distinct genotyping arrays, including five versions of the Illumina Bovine Low Marker Density (LMD) genotyping arrays (ranging from 9077 to 16,381 SNPs) and five versions of the EuroGenomics Medium Marker Density (MMD) arrays (ranging from 48,699 to 68,454 SNPs). The number of individuals per array are reported in Additional file 1: Table S1 (only individuals with a call rate > 0.90 were selected for this study).

The phenotypes included four linear classification scores, that assessed the muscular development of the shoulder, top and buttocks (rear and side view) of the animals on a scale of 0 to 50. To derive the overall score for muscular development, the individual scores were combined with different weights (1 for top and shoulder muscling, 2 for buttock muscling). These phenotypes, available for 16,508 of the cows, were corrected for fixed effects from the official genetic evaluation as described in Gualdrón Duarte et al. (2023).Two reference panels of bulls were available for genotype imputation, including a group of 717 AI bulls genotyped with the Illumina BovineHD genotyping array and whole-genome sequence data from 230 bulls. Details of the bioinformatic analysis of the sequence data, including read mapping and variant calling and filtering, can be found in Gualdrón Duarte et al. (2023) The final Variant Calling file (VCF) from the 230 sequenced bulls included 15,332,952 variants (12,830,339 SNPs and 2,502,613 indels). From these, we selected only bi-allelic autosomal variants.

5.3.2 Genotype imputation

A multi-step genotype imputation procedure was applied. First, SNP filtering was performed separately for each LMD and MMD array. SNPs with low call rate (<0.95), with minor allele frequency (MAF) < 0.01 or with significant deviations from Hardy-Weinberg proportions ($p > 0.001$) were filtered out. We first performed imputation from the LMD arrays to the MMD level, one array at a time. The MMD panel consisted of all individuals genotyped on one of the five MMD arrays. After filtering SNPs based on the rules described above, 36,849 autosomal markers with MAF > 0.01 and a maximum of 5 Mendelian inconsistencies in duos or trios, common to these five arrays and also present on commercial Illumina MMD arrays were retained to define the reference MMD panel. The different LMD arrays had 7246, 7505, 7711, 7632 and 7775 SNPs in common with the reference MMD panel, respectively. The target and reference panels were then phased using ShapeIT4.2 (Delaneau et al., 2019) and imputation in the target panel was achieved using Minimac4 (Das et al., 2016). After imputation, we excluded markers with a MAF < 0.02 or an imputation accuracy below 0.90 (for each array separately), and MMD genotypes from all individuals were merged. After selecting markers shared with the High Marker

Density (HMD) reference panel, 31,112 were available for the second imputation step. We used the HMD reference panel previously prepared by Gualdrón Duarte et al. (2023), which contained 890 individuals (717 genotyped and 173 sequenced bulls) and 611,322 markers. The same imputation and filtering procedure was applied as in the first imputation step. Finally, the cows were imputed to the sequence levels using 578,934 markers and the reference panel of 230 sequenced bulls from the study of Gualdrón Duarte et al. (2023). After this last imputation step, we selected variants imputed with imputation accuracy > 0.90, MAF > 0.01 and segregating according to HWE rules (p > 0.001), leaving 11,280,414 autosomal bi-allelic SNPs and indels for subsequent analyses.

5.3.3 Genomic prediction models

General genomic prediction models. We first describe models that don't use functional annotation, including GBLUP and three Bayesian models. In this case, the annotation information can be used, for example, to pre-select the variants to be included in the genomic prediction models.

In the GBLUP model, phenotypes are modelled as:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{g} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is the vector of individual phenotypes, $\mathbf{1}$ is a vector of 1's, μ is the mean effect, $\boldsymbol{g}$ is the vector of individual polygenic terms, and $\boldsymbol{e}$ is the vector of individual independent random error terms, normally distributed, $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$ where $\boldsymbol{I}$ is the identity matrix and σ_e^2 is the residual variance. The polygenic effects are normally distributed, $\boldsymbol{g} \sim N(0, \boldsymbol{G}\sigma_g^2)$ where $\boldsymbol{G}$ is the GRM and σ_g^2 is the variance of polygenic effects. The GRM can be computed using the matrix $\mathbf{Z}$ of centered genotypes, corresponding to the first rules proposed by VanRaden (2008) and assuming that the distribution of SNP effect does not depend on allele frequencies:

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{\sum_{j=1}^{N} 2f_j(1-f_j)},$$

where f_j is the allele frequency at marker j and N is the number of markers. Alternatively, the GRM can be obtained using the matrix $\mathbf{X}$ of centered and scaled (or "standardized") genotypes as described in Yang et al. (2011a):

$$\mathbf{G} = \frac{\mathbf{X}\mathbf{X}'}{N}.$$

In this case, rare alleles have larger effects and all variants contribute equally to the genetic variance. We will use GBLUP-C and GBLUP-S to refer to GBLUP with centered and standardized genotypes, respectively. The GRM and GBLUP prediction calculations were performed using LDAK

(Speed et al., 2012). For the GBLUP model, the variance components were estimated using a restricted maximum likelihood (REML).

For the Bayesian models, phenotypes are described as:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z\beta} + \boldsymbol{e},$$

where **β** is the vector of SNP effects. The models can be applied with centered or standardized (replacing $\boldsymbol{Z}$ by $\boldsymbol{X}$) genotypes. A key difference between the Bayesian models is the distribution of SNPs effects β_j. In BayesCπ (Habier et al., 2011), a fraction π of SNPs have a null effect:

$$\beta_j \sim \pi\ \delta_0 + (1-\pi)N(0, \sigma_\beta^2),$$

where δ_0 is a discrete probability mass at 0. The proportion of SNPs with zero effect (π) and the common variance of SNP effects σ_β^2 are estimated from the data. BayesCπ was run using the GCTB software (Zeng et al., 2018) with default settings.

In the Bayesian Sparse Linear Mixed Model (BSLMM) (Zhou et al., 2013), SNP effects are also distributed as a mixture of two distributions, with all SNPs having at least a small effect and a few SNPs having a large effect:

$$\beta_j \sim \pi\ N(0, \sigma_a^2 + \sigma_b^2) + (1-\pi)N(0, \sigma_b^2),$$

where σ_b^2 is the variance of small effects, σ_a^2 is the additional variance associated with large effects. The parameter π is now the proportion of SNPs with large effects. As in BayesCπ, the parameters are estimated. The model is implemented by modelling a polygenic term and using the associated GRM. BSLMM was run using the GEMMA software (Zhou et al., 2013) with default settings.

Finally, in BayesR (Erbe et al., 2012), the SNP effects are sampled from a mixture of four distributions:

$$\beta_j \sim \pi_1\delta_0 + \pi_2 N\left(0, 10^{-4}\sigma_g^2\right) + \pi_3 N\left(0, 10^{-3}\sigma_g^2\right) + \pi_4 N\left(0, 10^{-2}\sigma_g^2\right),$$

where π_1, π_2, π_3 and π_4 are the proportions of SNPs in the four categories. Where π_1 is the proportion of SNPs with null effects and π_4 is the proportion of SNPs with the largest effects, corresponding to one percent of the polygenic variance. The variances associated with each category are predetermined as fixed proportions of the polygenic variance which is estimated from the data as the mixture proportions. BayesR was run using the GMRM software (Patxot et al., 2021) without annotation (see below for more information).

Genomic predictions models exploiting prior biological information. Two methods were applied to perform whole genome predictions using directly information from functional annotations.

For this purpose, each SNP is assigned to one of the annotation groups, referred to as genomic features (GFs) to align with the terminology used in the literature (Edwards et al., 2016; Sørensen et al., 2014), described in the next section. First, we used a GBLUP, in which a distinct polygenic term is defined for each GF. The principle is similar to the MultiBLUP model described by Speed and Balding (2014) and the GFBLUP which fits most often a single GF at a time. We therefore call this model a Multiple Genomic Feature BLUP (MGFBLUP), and implement it as follows:

$$\boldsymbol{y} = \mathbf{1}\mu + \sum_{s=1}^{S} \boldsymbol{g}_s + \boldsymbol{e},$$

where $\boldsymbol{g}_s$ is the vector of individual polygenic terms associated to GF *s*, *S* is the total number of fitted GF. Each polygenic component is normally distributed, $\boldsymbol{g}_s \sim N(0, \boldsymbol{G}_s\sigma_s^2)$ where $\boldsymbol{G}_s$ is the GRM computed using the variants present in GF *s* and σ_s^2 is the variance of polygenic effects from the GF. As for the GBLUP, centered or standardized GRM can be used (MGFBLUP-C versus MGFBLUP-S). The genetic parameters, including the variances associated with each GF and the residual variance, were estimated using a REML approach as implemented in LDAK (Speed et al., 2012).

The second approach is a Bayesian grouped mixture of regressions model (GMRM), also called BayesRR-RC (Patxot et al., 2021) and derived from BayesR (Erbe et al., 2012) and BayesRC (MacLeod et al., 2016). In this model, phenotypes are described as:

$$\boldsymbol{y} = \mathbf{1}\mu + \sum_{s=1}^{S} \boldsymbol{X}_s \boldsymbol{\beta}_s + \boldsymbol{e},$$

where $\boldsymbol{X}_s$ is the matrix of centered and scaled genotypes for markers in GF *s* and $\boldsymbol{\beta}_s$ is the vector of marker effects for GF *s*, that are modelled as a mixture of null effects (spike probability at zero) and Gaussian distributions:

$$\beta_{s_j} \sim \pi_{0_s}\delta_0 + \pi_{1_s}N(0, \sigma_{1_s}^2) + \pi_{2_s}N(0, \sigma_{2_s}^2) + \ldots + \pi_{L_s}N(0, \sigma_{L_s}^2),$$

where *j* is the marker index, *L* is the number of Gaussian distributions in the mixture, $\{\pi_{0_s}, \pi_{1_s}, \pi_{2_s}, \ldots, \pi_{L_s}\}$ are the mixture proportions for GF *s*, $\{\sigma_{1_s}^2, \sigma_{2_s}^2, \ldots, \sigma_{L_s}^2\}$ are the mixture variances for GF *s*, proportional to σ_s^2, the variance explained by the GF. Here, *L* is equal to 3, with variances $\sigma_{l_s}^2$ equal to 0.0001, 0.001 and 0.01 σ_s^2, respectively. The hyper-parameters vary for variants from different GFs, and the variances σ_s^2 are estimated from the data. This model was run using the GMRM software (Patxot et al., 2021) with a Gibbs sampling scheme for 5,000 iterations with a burn-in period of 2,000 iterations. This setting corresponds to the values used by Patxot et al. (2021) and Orliac et al. (2022). When BayesRR-RC is used without annotation, we will refer to it as a BayesR model.

With these two approaches exploiting functional annotation, it is possible to define two parameters related to the contribution of a category to heritability and the relative size of effects in a category. First, the proportion of genetic variance explained by a category, also called percentage of heritability or %SNP heritability (Gusev et al., 2014), estimated as σ_s^2 divided by σ_g^2. Second, the enrichment level in category *i*, defined by Gusev et al. (2014) as the percentage of heritability in category *i* divided by the proportion of variants in the same category.

5.3.4 Annotation

We considered protein-coding variants to be those that alter the protein (e.g. change in amino acid sequence, truncations, alternative splice sites). To identify such variants, we ran Variant Effect Prediction (VEP) v95.0 (McLaren et al., 2016) on our VCF file. The most common coding consequences were missense, splice site (donor and acceptor), frameshift and stop-gain variants. VEP also provides the predicted effect from the variants, which is MODERATE or HIGH for coding variants and MODIFIER or LOW for other variants. Therefore, this first category contains all the variants with the highest predicted impacts.

We used three sources of information to identify putative regulatory variants. EQTLs provide the most direct evidence, as these variants present significant association with expression levels. Therefore, we extracted all cis-eQTLs from the cattle Genotype-Tissue Expression atlas (cGTEx) data base (Liu et al., 2022). For each eQTL we selected the lead SNP. This resulted in the selection of 22,817 eQTLs, including 4,889 eQTLs identified in muscle. In addition, variants located in open chromatin regions represent potential regulatory variants. Therefore, we used the catalogue of regulatory elements detected by the assay for transposase accessible chromatin using sequencing (ATAC-Seq) generated by Yuan et al. (2023). This organism-wide catalogue contains 976,813 cis-acting regulatory elements in 68 bovine tissues types. Variants located in these peaks represented 10% of the genome space. Finally, regulatory elements identified by Kern et al. (2021) in eight tissues, including muscle, were also considered as possible regulatory variants. These regulatory elements were identified thanks to epigenetic data for four histone modifications and one DNA binding protein (CTCF), and by applying ChromHMM (Ernst and Kellis, 2012) to predict genome-wide chromatin states in each tissue. Among the identified states, we selected active regulatory element states, including “CTCF / Active TSS”, “Active TSS”, “CTCF / promoters”, “Active promoters”, “CTCF / enhancers” and “Active enhancers”, where TSS stands for transcription start sites. All of these active marks are associated with the co-occurrence of at least two histone modifications and/or CTCF binding, and broad marks (e.g. associated only with the histone modification H3K27me3) were excluded.

We relied on the General Transfer Format (GTF) file of the bovine genome assembly available from Ensembl (v105) to classify the remaining variants. First, TSS and transcription termination sites (TTS) were obtained using Homer (Heinz et al., 2010) and all transcripts from the genes. Upstream and downstream regions were then defined as 1 kb upstream and downstream of the TSS and TTS,

respectively. Variants were then classified into three additional groups including ”Exon-associated elements” (encompassing exons and neighboring regions such as untranslated regions (UTRs) and regions upstream or downstream of genes), intronic regions, and intergenic regions corresponding to the remaining unannotated regions. Note that the Exon-associated elements contain only non-coding variants (e.g. synonymous variants) and putative regulatory regions that were not detected by the functional assays (i.e. not in the eQTL or regulatory element lists). There is in fact a hierarchy between the defined groups; if a variant can be associated with more than one group, we have chosen the group with the highest expected effect. The ranking of the groups, from most to least impactful, includes coding variants, eQTLs, regulatory elements (identified by ATAC-Seq or with epigenetic data), exon-associated elements, intronic and intergenic variants.

5.3.5 Experimental design

To assess the prediction accuracy of different models, we performed a cross-validation analysis and divided our data set into a reference and a target population corresponding to 13,461 and 3047 cows born before and after 1st January 2019, respectively. We then applied the different models with different marker panels and using different annotation groups. In most cases, we limited the number of annotation groups to 8 because more groups could lead to convergence problems with REML (or to null variances). Accuracy was obtained as the correlation between genomic estimated breeding values (GEBV) and trait deviations, while reliability was obtained as the squared correlation divided by the heritability of the trait.

As done by Meuwissen et al. (2024), we used a bootstrapping strategy to evaluate the significance of the difference in reliability between different methods or when using different marker panels. We created a table with the GEBVs of the > 3000 target individuals (rows) for all tested methods (columns). We then sampled the validation individuals with replacement 10,000 times and estimated the correlation between GEBVs and trait deviations, and estimated the reliability or reliability difference for each sample. The 2.5th and 97.5th quantiles were used to define the confidence intervals. Differences were considered significant if one method was higher in 97.5% or more of the samples.

Genomic prediction using whole-genome sequence data. We started by using all 11,280,414 variants available at the sequence level without annotation and ran centered and standardized GBLUP and BayesR. BayesCπ and BSLMM were not run on the full sequence for computational reasons. Next, we defined a first functional annotation model with eight groups (FAN1): coding variants, eQTLs, variants in regulatory elements identified by both ATAC-Seq and with epigenetic data, variants in regulatory elements detected with epigenetic data only, variants in regulatory elements detected by ATAC-Seq only, exon-associated elements, intronic regions and intergenic regions. With the second annotation model, we investigated whether separating regulatory variants identified in muscle from those identified only in other tissues improved prediction. In this case, the putative regulatory variants group contained variants in regulatory elements identified by ATAC-Seq or with epigenetic data. This

resulted in the following eight annotation groups (FAN2): coding variants, muscle eQTLs, other eQTLs, variants in muscle regulatory elements, variants in other regulatory elements, exon-associated elements, intronic and intergenic regions. In addition, we tested whether a stratification model based on LD and MAF (LDMS) improved prediction accuracy, as Orliac et al. (2022) have shown that these groups are important to include. We defined three MAF categories (0.01 < 0.05; 0.05 - 0.10; 0.10 - 0.50) and four LD-based categories (defined based on the LD score quartiles). These LD scores were calculated using GCTA (Yang et al., 2011a). We also combined the LDMS and FAN1 models, resulting in 8 x 12 groups. Note that this last model was only run with GMRM (Patxot et al., 2021), as we previously observed that the REML approach often had convergence problems when fitting a model with 12 or more groups (Yuan et al., 2024). These different models are described in Table 5.1, including the (functional) groups fitted in the models, and the number of variants per group.

Finally, the BayesRR-RC model was run twice with the FAN1 model to assess the variability in heritability partitioning across functional classes and prediction accuracy. In order to identify possible confounding between classes, we computed the correlations between the variances estimated in different iterations.

Use of biological information to pre-select markers. We then used the functional annotation groups to pre-select variants from the WGS data, as has been done in several studies (MacLeod et al., 2016; Xiang et al., 2021b, 2021a, 2019b). This is an indirect approach to include biological information with models that can't incorporate it directly (GBLUP, BayesCπ, BayesR and BSLMM), and amounts to assume that variants in unselected categories have a null effect. More importantly, it allows the data set and computational costs to be reduced, thus allowing the use of other models such as BayesCπ and BSLMM. Here, we selected a large subset of markers. This was done to include a high proportion of coding and regulatory variants and still capture the majority of sequence-level variants through LD.

Our first selection (Panel FUN1) included markers from the MMD panel currently used in the genomic evaluation and all coding and putative regulatory variants, including eQTLs and variants in regulatory elements detected by ATAC-Seq or with epigenetic data, resulting in a selection of 1,715,587 variants. We also defined a second panel (Panel FUN2) with fewer markers. It was generated using the same rules as above, except that putative regulatory elements were identified only based on the open chromatin regions defined by Yuan et al (2023). This amounts to using only one catalogue of putative regulatory elements and resulted in a panel with 1,284,915 markers. Similarly, we defined Panel-FUN3 using instead the catalogue of regulatory elements from Kern et al. (2021) and obtained 863,615 markers. For comparison, we generated other panels obtained by performing LD pruning (based on the r^2 measure) with thresholds of 0.99, 0.98, 0.95, 0.90 and 0.80, resulting in selection of 1,899,123, 1,708,694, 1,436,932, 1,203,927 and 923,968 variants, respectively (Panels LD99 to LD80). In addition, we selected all markers present on commercial bovine genotyping arrays extracted from the SNPchiMp data base (Nicolazzi et al., 2014), resulting in 868,195 polymorphic SNPs (Panel ARRAY). Table 5.2 summarizes all the defined panels, their size and how the markers were selected.

For all panels, we first ran models without annotation, including centered and standardized GBLUP, BayesCπ, BSLMM and BayesR. For panels FUN1, FUN2 and FUN3, we ran BayesRR-RC and MGFBLUP with similar functional groups as for the sequence data, but adapted as some categories were removed from the data. FAN1 now contained four groups (MMD markers, coding variants, eQTLs, variants in regulatory elements) while FAN2 still contained six groups (MMD markers, coding variants, muscle eQTLs, other eQTLs, variants in muscle regulatory elements, variants in other regulatory elements).

Table 5.1. Description of the different annotation models and their respective categories.

Model	Annotation Group	Number of variants	Proportion in the genome
FAN1: eight functional annotation groups allowing distinct effect sizes for coding and regulatory variants	Coding variants	41,866	0.37%
	eQTLs	31,521	0.28%
	Regulatory elements detected by ATAC-Seq	855,103	7.58%
	Regulatory elements detected with epigenetic data	431,616	3.83%
	Regulatory elements detected with both techniques	333,877	2.96%
	Exon-associated elements	732,544	6.49%
	Intronic	2,994,362	26.54%
	Intergenic	5,859,525	51.94%
FAN2: eight functional annotation groups similar to FAN1 but with specific categories for regulatory elements detected in muscle	Coding variants	41,866	0.37%
	eQTLs detected in muscle	4,761	0.04%
	eQTLs detected in other tissues	26,760	0.24%
	Regulatory elements detected in muscle	80,378	0.71%
	Regulatory elements detected in other tissues	1,540,218	13.65%
	Exon-associated elements	732,544	6.49%
	Intronic	2,994,362	26.54%
	Intergenic	5,859,525	51.94%
LDMS: 12 groups based on the combination of four LD groups based on LD score and three MS groups based on MAF values	LD: Four equal groups based on LD score quartiles	2,820,104	25.00%
	MS: Minor allele frequency between 0.01 and 0.05	2,193,621	19.45%
	MS: Minor allele frequency between 0.05 and 0.10	1,663,801	14.75%
	MS: Minor allele frequency between 0.10 and 0.50	7,422,992	65.80%
LDMS x FAN1: interaction between FAN1 and LDMS model	96 groups based on the combination of the 12 LDMS groups and the 8 FAN1 groups	From 29 to 1,446,999	From 0.00% to 12.83%

Table 5.2. Description of the different selected marker panels.

Panel	Selection criteria	Number of variants
WGS	All whole-genome sequence variants	11,280,414
FUN1	Coding variants, eQTLs, variants in regulatory elements detected by ATAC-Seq or epigenetic data and markers on MMD array	1,721,775
FUN2	Coding variants, eQTLs, variants in regulatory elements detected by ATAC-Seq only and markers on MMD array	1,292,091
FUN3	Coding variants, eQTLs, variants in regulatory elements detected by epigenetic data only and markers on MMD array	870,858
LD99	Selection based on LD pruning with a threshold of $r^2 > 0.99$	1,899,123
LD98	Selection based on LD pruning with a threshold of $r^2 > 0.98$	1,708,694
LD95	Selection based on LD pruning with a threshold of $r^2 > 0.95$	1,436,932
LD90	Selection based on LD pruning with a threshold of $r^2 > 0.90$	1,203,927
LD80	Selection based on LD pruning with a threshold of $r^2 > 0.80$	923,968
ARRAY	Variants from different commercial bovine genotyping arrays	868,195

5.4 Results

5.4.1 Genomic prediction models using all sequence-level variants

Reliabilities obtained with GBLUP-C using the MMD panels were 0.792, 0.674, 0.686, 0.750 and 0.705 for shoulder, top, buttock (side view), buttock (rear view) and overall muscling, respectively. As the GBLUP-C is the approach currently used in genomic evaluation, we used these as the baseline or reference values and presented results from models using the full sequence data as the difference from these baseline values (Figure 5.1). We first evaluated models using the 11 million imputed variants (Figure 5.1; Additional File 2: Table S8). This was only possible with GBLUP (without annotation) and MGFBLUP (fitting one polygenic term per annotation group) approaches using either centered or standardized genotypes, and with the GMRM program fitting the BayesR (without annotation) and BayesRR-RC (with annotation) models. Compared to MMD arrays, the use of WGS data consistently resulted in higher accuracy for the three annotation-free models and for all traits (Additional File 2: Table S8). With the GBLUP, the use of centered genotypes (GBLUP-C) gave better results (+1.6% reliability on average) than standardized genotypes (GBLUP-S) (+0.6% on average). Prediction accuracies achieved with BayesR (+1.8% on average), implemented using standardized genotypes, were systematically higher than those obtained with GBLUP-S, suggesting advantages of the Bayesian approach. However, the superiority was less pronounced when compared with GBLUP-C. Although these trends were consistent across traits, only a few of these differences were significant (Additional File 3: Figure S1).

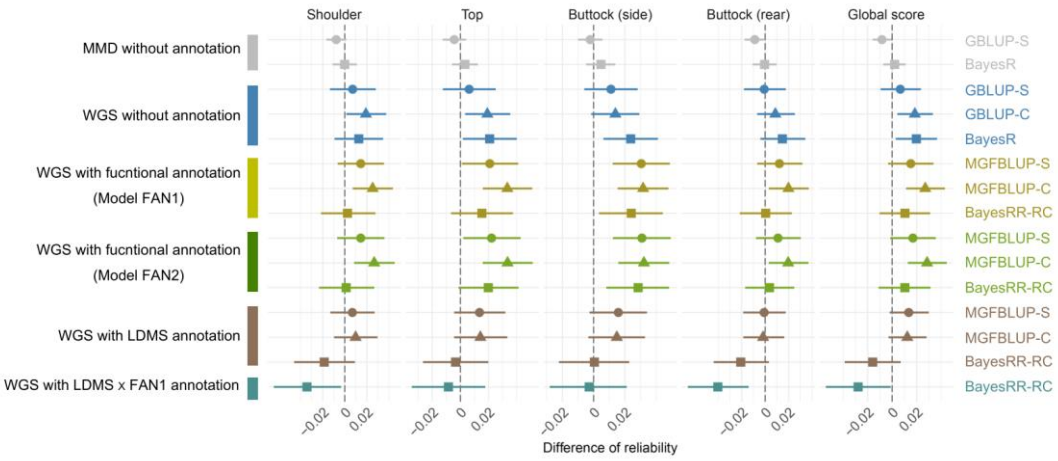


Figure 5.1. Gain of reliability obtained when using whole-genome sequence data, with or without functional annotation. The gain in reliability compared to a GBLUP-C model with a medium marker density (MMD array). GBLUP and BayesR correspond to models without functional annotation, while the Multiple Genomic Feature BLUP (MGFBLUP) and BayesRR-RC refer to extensions of these models that make use of functional annotation. For the GBLUP and MGFBLUP models, the extensions '-C' and '-S' indicate whether the used GRMs were constructed with centered and standardized genotypes, respectively. Genomic predictions were performed using different panels (MMD and whole-genome sequence – WGS) and models including models without annotation, two models incorporating functional annotation (FAN1 & FAN2), a model based on LD and MAF stratification (LDMS), and a combination of LDMS and FAN1 models. Further details of these models are provided in Table 5.1. Error bars indicate the 95% confidence interval of the bootstrapped differences.

Compared to GBLUP, the reliabilities obtained using functional annotation with the MGFBLUP approach were higher (Figure 5.1), although the differences were rarely significant (Additional File 3: Figure S1). For example, with centered genotypes, the MGFBLUP reliabilities were on average +1.2% higher for the two functional annotation models tested (FAN1 and FAN2). The opposite trend was observed for the Bayesian models, the reliabilities of BayesRR-RC with the FAN1 and FAN2 models being on average –0.8% and –0.6% lower than those obtained with BayesR (without annotation). When groups were defined based on LD and MAF (LDMS models), lower reliabilities were obtained with MGFBLUP-C (-1.6% on average) and BayesRR-RC (-3.0% on average) compared to the corresponding models without annotation. Reliabilities obtained with BayesRR-RC were particularly low when fitting 96 groups with the LDMS x FAN1 model (-4.1% on average). Overall, of all the methods tested, the MGFBLUP-C model with functional annotation achieved the highest accuracies for each of the traits.

The proportion of variance allocated to different functional categories and the enrichment levels of variants with different annotations allow a better understanding of how the models use the functional information. For example, estimated parameters from the first model (FAN1) using functional annotation (see Table 3 for average values and Additional File 1: TableS2-6 for values per trait) showed that coding variants and eQTLs had larger effects per SNP on average (average enrichment levels above 15-fold and 20-fold, respectively), followed by variants in putative regulatory elements with average enrichment levels ranging from 1.7 to 3.9-fold depending on the method (Table 5.1). With MGFBLUP models, eQTLs had even larger effects than coding variants (e.g. 16.1-fold versus 70.5-fold when using centered genotypes). Nevertheless, these relatively small groups (each containing less than 0.4% of the variants) together accounted for only 10-25% of the genetic variance, whereas intergenic and intronic variants still accounted for a large proportion of the genetic variance (about 40%), as together they represent more than 75% of the variants in our data set. With MGFBLUP and the second annotation model (FAN2), variants associated with eQTLs or regulatory elements detected in muscle had higher enrichment levels than variants in the same elements detected in other tissues (e.g. muscle eQTL enrichment levels were on average higher than 100-fold), whereas the opposite was observed for eQTLs with BayesRR-RC (Table 5.3). Some unexpected results were observed, such as reduced enrichment levels for coding variants with the FAN2 model and BayesRR-RC, or a null variance associated with the category of exon-associated elements when estimated with MGFBLUP (Table 5.3). Such results indicate that parameters can be difficult to estimate (see also Yuan et al., 2024) and that enrichment levels used in predictions may not always reflect true biological enrichment levels. The variation in estimated parameters across traits (Additional File1: TableS2-6; Figure 5.2A) confirmed this technical difficulty. Interestingly, this variation had little effect on the relative performance of the different models. To further understand aspects of convergence with BayesRR-RC, we ran an additional independent chain for the FAN1 model (including more iterations), generated some diagnostic plots, and assessed the level of confounding between categories from their correlation across iterations (Figure 5.2A-D). Estimated genetic variances for different functional categories showed differences across

independent runs (Figure 5.2A), particularly for coding variants, while diagnostic plots suggested that convergence may not have been achieved for all parameters (Figure 5.2C), possibly due to confounding between some parameters (e.g. between eQTLs and variants in regulatory elements detected by ATAC-Seq categories; Figure 5.2D). Despite these differences in estimated enrichment levels, similar prediction accuracies were obtained with the two independent chains and when more iterations were run (Figure 5.2B).

Table 5.3. Average estimated %SNP heritability (proportion of genetic variance explained by a category) and enrichment levels (relative variant effect size per category) for different functional categories with the two functional annotation models, FAN1 and FAN2. MGFBLUP models were applied with GRMs constructed with either centered or standardized genotypes.

	Annotation Group (compartment where variants are located)	%SNP heritability			Enrichment		
		MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC	MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC
FAN1	Coding variants	5.98	7.65	7.99	16.11	20.61	21.54
	eQTLs	14.91	17.33	4.36	70.50	81.93	20.59
	Regulatory elements detected with both techniques	7.87	8.21	11.59	2.65	2.77	3.91
	Regulatory elements detected by ATAC-Seq	20.48	12.96	13.13	2.70	1.71	1.73
	Regulatory elements detected with epigenetic data	8.74	6.89	11.74	2.28	1.80	3.06
	Exon-associated elements	0.22	0.07	7.09	0.03	0.01	1.09
	Intronic regions	26.87	31.11	21.60	1.01	1.17	0.81
	Intergenic regions	14.94	15.78	22.49	0.29	0.30	0.43
FAN2	Coding variants	5.88	8.03	3.45	15.84	21.64	9.28
	eQTLs detected in muscle	4.57	4.71	0.24	108.22	111.56	5.71
	eQTLs detected in other tissues	11.28	12.72	5.82	57.89	65.27	29.86
	Regulatory elements detected in muscle	9.35	9.52	13.69	13.11	13.35	19.19
	Regulatory elements detected in other tissues	24.46	18.14	19.99	1.79	1.33	1.46
	Exon-associated elements	0.49	0.10	12.26	0.08	0.02	1.89
	Intronic regions	27.15	30.24	22.03	1.02	1.14	0.83
	Intergenic regions	16.83	16.55	22.53	0.32	0.32	0.43

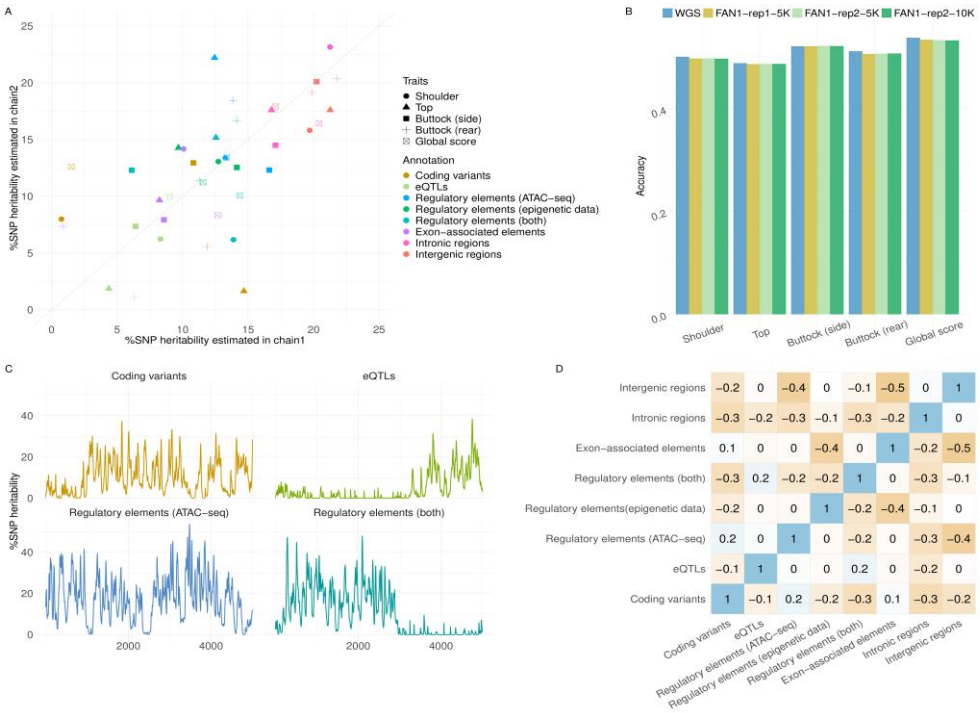


Figure 5.2. Comparison of results of BayesRR-RC with two independent chains. A) Percentage heritability (%SNP heritability) of different functional categories estimated in two independent runs. B) Accuracy of GS in the two independent chains for the FAN1 model compared to accuracy with WGS (without annotation); results are also reported with 10,000 iterations for the second chain (instead of 5,000). C) Evolution of estimated parameters over iterations for four functional categories. D) Correlations between estimated parameters for different functional categories in different iterations (when estimated for buttock – side view).

5.4.2 Genomic prediction models using subsets of the sequence data

We compared different strategies for selecting large subsets of the sequence data, large enough to still capture the full sequence level while allowing the use of additional, more computationally demanding software, including GCTB for BayesCπ (Habier et al., 2011) and GEMMA for BSLMM (Zhou et al., 2013). For all methods, the highest accuracy was achieved in the vast majority of cases with an LD pruning level of $r^2 > 0.99$ (1.9 M variants) (Figure 5.3A; Additional File 2: Table S9). Accuracy was even higher than with full sequence data (for BayesR and GBLUP approaches). Reliabilities decreased only slightly when stronger LD pruning was applied and the number of variants was further reduced (< 1% on average). For traits such as buttock muscling (side view), the reliabilities remained almost the same even when using an LD pruning level of $r^2 > 0.80$ (0.9 M variants), whereas the greatest reduction in reliability was observed for shoulder muscling. However, the differences between the largest and smallest marker panels were not always significant, depending on the method (Additional File 3: FigureS2).

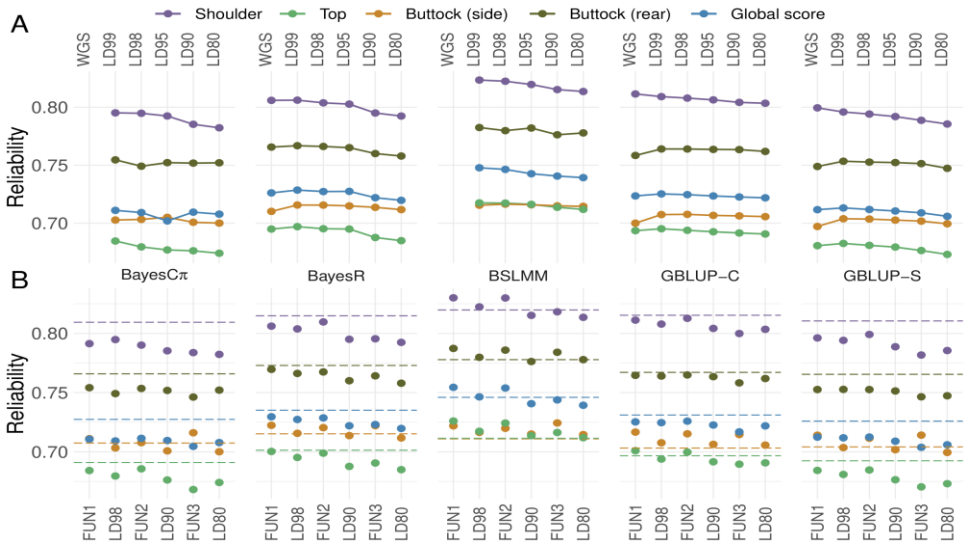


Figure 5.3. Comparison of reliability of different methods when using different marker panels. A) Comparison of the reliability of five tested methods using panels selected on the basis of LD pruning. The panels are whole-genome sequence (WGS - no pruning) and LD99, LD98, LD95, LD90 and LD80 obtained when pruning was applied with thresholds of $r^2 > 0.99$, 0.98, 0.95, 0.90 and 0.80, respectively. B) Reliability obtained using panels selected on the basis of functional annotation. The panels included coding variants, eQTL, variants in regulatory elements and markers from the medium density genotyping array. The FUN1 panel included variants in regulatory elements detected by ATAC-Seq or epigenetic data, while the FUN2 and FUN3 panels include only those detected by either ATAC-Seq or epigenetic data, respectively. The results were compared to those obtained with panels of equal size selected by LD pruning. The horizontal line corresponds to the reliability obtained with the ARRAY panel, obtained by selecting markers present on commercial bovine genotyping arrays, with approximately the same number of variants as the FUN3 panel. Further details on the different panels and their size are given in Table 5.2.

At all LD pruning levels, BSLMM was systematically the best approach (with a single exception, BayesR achieving slightly higher reliabilities for buttock side when using a LD pruning level of 0.99), while BayesCπ and GBLUP-S were often the least accurate (Figure 5.3A). The ranking between the BayesR and GBLUP approaches was consistent with that observed with the full sequence data, BayesR achieving on average higher reliability than both GBLUP approaches (for LD pruning levels of 0.95 or higher) and GBLUP-C being superior to GBLUP-S (Figure 5.3A). At the $r^2 > 99$ pruning level, the accuracies obtained with BSLMM were significantly higher than those achieved with BayesR, GBLUP-C, GBLUP-S, and BayesCπ for several traits (with the exception of buttock side with BayesR; Figure 5.3A; Additional File 3: Figure S3).

With BSLMM, the average number of variants with a large effect fitted in the model ranged from 22 to 138 (mean = 71.0) per trait (LD99), while with BayesR and BayesCπ, the average number of variants with a non-zero effect ranged from 10,990 to 11,507 (mean = 11,193) and 121,759 to 124,891 (mean = 135,818), respectively (LD99). With BayesR, the number of variants with a medium or large effects (0.001 and 0.01 x σ_g^2) were low, 3.8 and 1.5 on average, indicating that it was not able to exploit large effect

variants as well as BSLMM. Note that the definition of large effect variants is however not directly comparable between these two methods. With stronger LD pruning, the number of large effect variants increased slightly with BSLMM (85.8 and 105.4 on average for LD98 and LD80, respectively), while the proportion of non-zero effects (corresponding to π) remained relatively stable with BayesCπ (π remained close to 0.05). This was also the case for the number of medium and large effect variants with BayesRR-RC (e.g. 3.9 and 0.8 for LD80).

We then compared the reliabilities obtained with marker panels selected based on functional annotation (called FUN1-3 panels) with panels of equivalent size selected based on LD pruning (LD panels) or panels containing markers present on commercial bovine genotyping arrays (870K SNPs) (Figure 5.3B; Additional File 2: Table S9). For most methods, marker selection based on functional annotation resulted in slightly higher or equivalent accuracy than LD-based marker selection (Figure 5.3B), although the differences were almost never significant when compared at equal density (Additional File 3: Figure S4). In particular, the FUN1 and FUN2 panels were often more efficient than the corresponding LD panels. In BSLMM, the advantage of the FUN panels was observed for all traits and marker sizes. Higher reliabilities were systematically obtained with the ARRAY panel when using GBLUP-S or BayesCπ (except for buttock – side view), while for GBLUP-C and BayesR the accuracy was very close to that obtained with the two largest FUN panels. Importantly, the ARRAY panel was in most cases significantly superior to other panels of equivalent size for these four methods. With BSLMM, the use of the FUN1 and FUN2 panels resulted in higher reliabilities than the use of the ARRAY panel, while for the FUN3 panel the reliabilities obtained were either higher or equal to those obtained with the ARRAY panel (Figure 5.3B). Note that the use of FAN1 and FAN2 models incorporating functional annotation (BayesRR-RC and MGFBLUP) with the FUN1 and FUN2 panels did not improve the reliability of genomic prediction, while at best a slight improvement was observed when using the FUN3 panel (Figure 5.4; Additional File 2: Table S10).

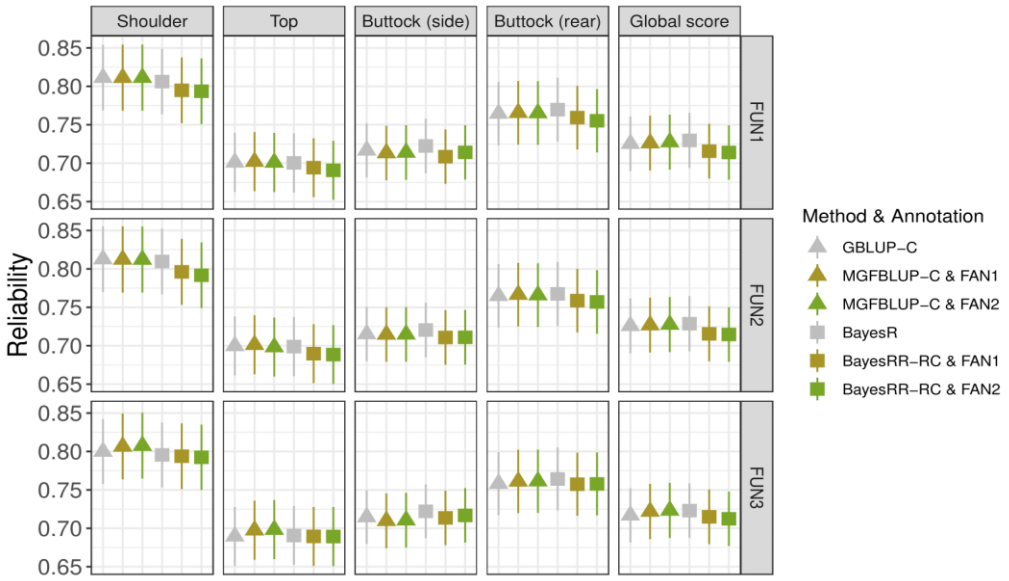


Figure 5.4. Reliability of models incorporating functional annotation applied to subsets of markers. Results were obtained using the three functional panels (FUN1-3) and applying GBLUP-C or BayesR without functional annotation or with the Multiple Genomic Feature BLUP (MGFBLUP) and BayesRR-RC models with the two functional models (FAN1 and FAN2 models). The FUN1 panel (top) includes variants in regulatory elements detected by ATAC-Seq or epigenetic data, while the FUN2 (middle) and FUN3 (bottom) panels include only those detected by either ATAC-Seq or epigenetic data, respectively. The FAN1 model has four groups, including coding variants, eQTLs, variants in regulatory elements and markers from the medium density genotyping array. In the FAN2 model, two additional categories are obtained by dividing eQTLs and variants in regulatory elements into those detected in muscle and those detected in other tissues. Details of the models and marker panels are given in Tables 5.1 and 5.2.

Overall, the highest reliabilities were obtained using BSLMM with variants selected based on their functional annotation, closely followed by BSLMM with LD panels. Compared with GBLUP-C using MMD markers, the reliabilities obtained with BSLMM for the FUN1 panel were significantly 0.052, 0.038, 0.036, 0.038 and 0.049 higher for top, shoulder, buttock (side and rear view) and overall muscling, respectively (Additional File 1: Table S7). Similar benefits were also obtained with the LD99 panel (Additional File 1: Table S7). With the FUN1 and FUN2 panels, the number of large effect variants ranged from 30 to 91 (mean = 63.9) and from 33 to 200 (mean = 86.2), respectively. The %SNP heritability associated with the large effect variants was 14.3% and 16.2% on average with the FUN1 and FUN2 panels, respectively. We investigated which regions contained variants with high posterior inclusion probabilities (PIP) when using the FUN2 panel (the reliabilities with the FUN2 panel were virtually identical to those with the FUN1 panel, but had the advantage of using fewer variants). We summed the PIP over 1 Mb windows to account for the possibility that different variants in LD might capture the same effect, and identified regions with a cumulative PIP > 0.5 (e.g., regions that had a large effect in more than half of the iterations) (e.g. Barbieri and Berger, 2004). We identified 5 to 12 regions per trait, many of which overlapped with the 15 large effect variants fine-mapped by Gualdrón Duarte et al. (2023).

5.5 Discussion

5.5.1 Increased prediction accuracy with whole-genome sequence data

In this study, we used whole-genome sequence data, with or without annotation, to improve the accuracy of genomic prediction of muscular development traits in BBC. Contrary to some previous studies (Ros-Freixedes et al., 2022; van Binsbergen et al., 2015; Veerkamp et al., 2016), the use of full-sequence data increased the reliability of breeding values compared to the use MMD arrays. Several factors may explain this improvement, including more reliable genotype imputation due to ever-expanding reference panels and improving imputation software, and the benefits of ever-larger reference populations for genomic prediction. Although the benefit of sequence data was systematic, it remained relatively modest (e.g., +1.8% reliability on average with BayesR) and was not always significant. This is consistent with our previous findings on simulated data (Druet et al., 2014b), which showed that the use of whole-genome sequence data allowed to increase reliability mainly when rare variants contributed to genetic variance and were accurately genotyped or imputed. Conversely, little gain was obtained when common variants accounted for the largest proportion of genetic variance, as observed here. The segregation of common variants with large effects on muscular development traits in BBC or height has been previously described (Gualdrón Duarte et al., 2023), while it remains difficult to study the importance of rare variants. Indeed, the imputation accuracy for rare variants remains low and we discarded variants with MAF < 0.01 or with low imputation accuracy. Therefore, we could not fully exploit the variation associated with rare variants.

5.5.2 Relative performance of prediction models

Our study was also informative about the differences between methods. With full sequence data, BayesR and BayesRR-RC were the only Bayesian models to individually fit all the SNPs that could be run on our cluster, thanks to their implementation in the GMRM software. BayesR had slightly higher reliabilities than GBLUP-S (+1.2% on average), while the differences were even smaller when compared to GBLUP-C (+0.2% on average). Indeed, in our study, the use of centered genotypes consistently performed better than standardized genotypes (commonly used in human studies). This ranking is in agreement with previous studies that have been carried out in BBC (Gualdrón Duarte et al., 2020) and in some other cattle breeds (e.g., Su et al., 2014). In agreement, we also observed that selecting common variants (ARRAY panel) with methods using standardized genotypes and giving more weight to rare alleles (e.g., GBLUP-S) increased the reliability. The relationship between MAF and effect size has previously been linked to ongoing selection in the populations analyzed, with the architecture corresponding to standardized genotypes (i.e. rare variants having larger effects) being associated with purifying selection, whereas centered genotypes (i.e. common variants accounting for large proportions of genetic variance) are

associated with directional selection (Zeng et al., 2018). Overall, these results therefore suggest directional selection for muscular development traits in BBC and support the use of centered genotypes. Bayesian models that allow for variants with large effects are expected to perform better than GBLUP when such variants contribute to the genetic architecture (Hayes et al., 2010). In the present case, however, the benefit remained modest. Although the reasons remain unknown, several elements may explain this observation. First, the fact that as the sample size increases, the priors have less influence on the estimated effects (i.e. the large effect variants are also well captured in the SNP-BLUP (e.g. Pocrnic et al., 2024)). Second, the frequency of some of the large effect variants previously identified in BBC and associated with recessive deleterious effects has decreased (see Additional File 3: Figure S5) as a result of selection against carrier bulls (i.e. there are fewer large effect variants in the target population). Thirdly, muscular development traits in BBC tend to be polygenic, more so than height for example (Gualdrón Duarte et al., 2023). We must also bear in mind that estimating the effects of more than 10 million variants simultaneously with 14K genotyped individuals remains a challenging task and may require further iterations. Importantly, the reliability of BayesRR-RC could be further increased when using centered genotypes in livestock species. When using a subset of SNPs, the BSLMM approach (Zhou et al., 2013) was consistently the best of all approaches, suggesting that a model combining a polygenic model with a few large effect variants is efficient. We also tested two other similar models fitting a polygenic term and a group of large effect variants, Bolt-LMM (Loh et al., 2015b) and BayesGC (Meuwissen et al., 2021), but these did not perform as well as BSLMM (data not shown). In fact, such approaches could also be fitted with the BayesR framework, as the number of mixtures and their relative variance can be modified. Note also that BayesR and various extensions, including BayesRC (MacLeod et al., 2016) or BayesRCO (Mollandin et al., 2022), have been implemented in different programs (Breen et al., 2022; Mollandin et al., 2022; Moser et al., 2015) and that some of these may achieve higher accuracy, due to differences in model assumptions, settings or genotype coding (centered versus standardized genotypes). For example, we previously observed smaller differences between BSLMM and BayesR on a smaller data set with fewer markers, although BSLMM was still more accurate on average (Gualdrón Duarte et al., 2020).

5.5.3 Benefits of using functional annotation

This study is one of the first to use full sequence data to perform genomic prediction using models that incorporate functional information from experimental data in livestock species and using a relatively large cohort of individuals. For Bayesian mixture models, this was possible thanks to the development of software such as GMRM. However, this approach did not result in the strong improvement in reliability reported for humans (Márquez-Luna et al., 2021; Patxot et al., 2021; Zheng et al., 2024). This may be due to differences in population structure and past selection history (see above). In livestock species, effective

population size is generally lower (e.g. Hayes et al., 2003; MacLeod et al., 2013), while LD extends at longer range (e.g. Gautier et al., 2007) and relatedness is higher, as evidenced by the level of inbreeding observed in BBC (Solé et al., 2017). This makes it more difficult to disentangle contributions from different functional classes, which are more confounded, as shown in the present study or previously in Yuan et al. (2024). As a result, the estimated enrichment levels are imprecise and the functional annotation is less optimally used. Another major difference is that the sample sizes used in human studies are much larger, providing more information to estimate different parameters. The genetic architecture also differs, as we have previously observed. While coding variants with large effects are rare and common variants have small effects and are generally regulatory in complex traits studied in humans (Eyre-Walker, 2010; Marouli et al., 2017; Zeng et al., 2018), common coding variants with large effects are regularly observed in livestock species (Gualdrón Duarte et al., 2023; Hayes et al., 2010; Kemper and Goddard, 2012). In addition, rare variants remain more difficult to exploit in livestock species due to low imputation accuracies and smaller reference panels. Finally, the amount and quality of functional information available in human studies is still much higher than in livestock species, allowing more genomic features to be fitted, such as in the so-called LD-baseline model including up to 53 groups. Better functional annotation could be achieved for the BBC breed by generating breed-specific regulatory variant catalogues using large numbers of individuals. Finally, additional categories could be considered, such as conservation scores, which have been shown to be relevant in both humans and livestock (Finucane et al., 2015; Xiang et al., 2019b).

When analyzing muscular development traits in BBC, the parameters estimated by models incorporating functional information were highly variable across traits and methods, and sometimes difficult to interpret biologically (e.g. null variance associated with some categories). These parameters should thus be interpreted with caution (Yuan et al., 2024), especially for small categories such as coding variants or eQTLs. In addition, the estimates are highly dependent on the definition of the different functional categories, which may differ between studies, making comparisons difficult. Although the enrichment levels of the different categories are ranked as expected, we do not recommend evaluating their absolute values.

In such settings, the MGFBLUP framework produced slightly higher accuracies than GBLUP, probably because these models had greater flexibility, whereas BayesRR-RC tended to decrease accuracy compared to BayesR. Even without annotation, this Bayesian approach already has great flexibility (i.e. allowing some variants to have larger effects), and there may be less benefit in adding more flexibility, especially when more parameters need to be estimated and the reference population is not large enough. We have observed the difficulties and challenges of estimating all these parameters and individual variant effect simultaneously with BayesRR-RC. A potential disadvantage of BayesRR-RC is that when annotation is used, the variance used to model the SNP effects is reduced. For example, without annotation, the largest

effects of coding variants are sampled from a distribution corresponding to 1% of the total genetic variance, whereas with annotation this becomes 1% of the variance explained by coding variants (i.e., if coding variants account for 10% of the SNP heritability, the effects are sampled from a distribution with a 10 times lower variance). This problem can be addressed in BayesRR-RC by fine-tuning the parameters of the model, as done by Orliac et al. (2022). Note also that the BayesRC model (MacLeod et al., 2016) uses the total genetic variance to model the SNP effects, which makes the model more robust to this problem and has the advantage of reducing the number of parameters to be estimated. Overall, in our setting, the accuracy of the models seems to result from their ability to capture the polygenic terms and the large effect variants, even with non-causal markers in LD with the causative variants, rather than from their ability to exploit the functional information.

5.5.4 Future directions

Using BSLMM on large subsets of variants, selected based on their functional annotation or LD pruning, gave the highest reliabilities and significantly improved genomic predictions. With BSLMM, about 50 to 100 variants with large effects were fitted in each iteration. Ideally, we should identify these 50-100 variants, or eventually a few more, and fit a model with a polygenic term and only these additional variants. It remains difficult to identify these variants with simple functional annotation because many of the functional classes are too large and lack specificity. For example, the number of coding variants and eQTLs is much larger (even if we target the 1,000 variants with the largest contribution to genetic variance), although they do not include all the large effect variants. Improved fine-mapping approaches using functional annotation are needed to identify more of these variants, as in general only a handful of causative variants are currently unambiguously identified. Further improvements in functional annotation are therefore needed, including experimental data in the most relevant tissues, experiments on relatively large samples of individuals from the same breeds, and finer annotation levels. For example, by defining categories that combine motifs of transcription factor that are specific to the correct pathway and their levels of conservation. It would also be important to be able to identify which synonymous variants, assumed to be neutral, have an effect on the traits of interest. Finally, additional work is required to better exploit rare variants, for which imputation accuracy remains low, in genomic prediction.

5.6 Conclusions

Compared to the GBLUP approach using medium maker density, as in the current genomic evaluation, the use of imputed whole-genome sequence data allowed to increase the reliability of genomic predictions for muscular development traits in BBC (+1.8% on average with the best method). Selection of subsets of markers based on functional annotation or LD pruning, allowed equivalent accuracy to be

achieved at lower computational cost, allowing more methods to be applied. Overall, a strategy using a large panel of pre-selected functional variants, including coding variants, eQTLs and variants in regulatory elements, with a Bayesian model fitting a polygenic term combined with fewer than 200 large effect variants achieved the highest accuracies (+4.2% on average). Therefore, fine-mapping of these large effect variants may prove effective in improving genomic prediction accuracy. Models directly incorporating functional annotation only slightly improved reliability at best. This suggests that better annotation categories should be used than in the present study, and that further efforts are needed to improve functional annotation in BBC. In addition, more work is needed to exploit the genetic variance associated with rare variants, which remain difficult to impute accurately.

5.7 Acknowledgements

The authors acknowledge the Walloon Breeders Association for providing the data. Tom Druet is Research Director from the F.R.S.-FNRS. We used the supercomputing facilities of the "Consortium d'Equipements en Calcul Intensif en Fédération Wallonie-Bruxelles" (CECI), funded by the F.R.S.-FNRS.

5.8 Supplementary tables

Table S1. Number of individuals and markers per genotyping array.

Genotyping array	Number of individuals	Number of variants	Marker density
LMD1	2,467	9,077	Low
LMD2	570	12,023	Low
LMD3	1,511	12,638	Low
LMD4	8,165	15,749	Low
LMD5	583	16,381	Low
MMD1	4,930	48,699	Medium
MMD2	13,036	54,748	Medium
MMD3	1,843	62,227	Medium
MMD4	1,690	68,454	Medium
MMD5	1,394	49,229	Medium

Table S2. Estimated %SNP heritability (proportion of genetic variance explained by a category) and enrichment levels (relative variant effect size per category) for different functional categories with two annotation models, FAN1 and FAN2, when applied to shoulder muscling. MGFBLUP models were applied with GRMs constructed with either centered or standardized genotypes, respectively. The annotation categories are described in the Methods section.

	Annotation Group (compartment where variants are located)	%SNP heritability			Enrichment		
		MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC	MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC
FAN1	Coding variants	2.36	5.06	0.01	6.35	13.63	0.03
	eQTLs	17.53	20.07	5.76	82.88	94.85	27.25
	Regulatory elements detected with both techniques	0.79	7.17	14.19	0.27	2.42	4.78
	Regulatory elements detected by ATAC-Seq	23.19	17.18	10.36	3.06	2.26	1.36
	Regulatory elements detected with epigenetic data	11.41	0.02	12.81	2.98	0	3.34
	Exon-associated elements	0.06	0.04	9.1	0.01	0.01	1.4
	Intronic regions	28.54	34.47	24.14	1.07	1.3	0.91
	Intergenic regions	16.12	15.99	23.63	0.31	0.31	0.45
FAN2	Coding variants	1.28	3.98	0	3.46	10.73	0
	eQTLs detected in muscle	7.07	8.01	0.69	167.43	189.87	16.36
	eQTLs detected in other tissues	13.45	14.13	10.7	69.03	72.52	54.92
	Regulatory elements detected in muscle	8.63	11.25	14.17	12.09	15.77	19.86
	Regulatory elements detected in other tissues	18.94	7.11	21.74	1.39	0.52	1.59
	Exon-associated elements	0.07	0.08	13.97	0.01	0.01	2.15
	Intronic regions	30.35	35.51	21.11	1.14	1.34	0.79
	Intergenic regions	20.21	19.93	17.63	0.39	0.38	0.34

Table S3. Estimated %SNP heritability (proportion of genetic variance explained by a category) and enrichment levels (relative variant effect size per category) for different functional categories with two annotation models, FAN1 and FAN2, when applied to top muscling. MGFBLUP models were applied with GRMs constructed with either centered or standardized genotypes, respectively. The annotation categories are described in the Methods section.

	Annotation Group (compartment where variants are located)	%SNP heritability			Enrichment		
		MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC	MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC
FAN1	Coding variants	10.88	13.57	16.76	29.32	36.58	45.17
	eQTLs	15.81	18.13	1.58	74.71	85.69	7.46
	Regulatory elements detected with both techniques	1.91	0.00	12.91	0.64	0.00	4.35
	Regulatory elements detected by ATAC-Seq	21.32	16.68	9.27	2.81	2.20	1.22
	Regulatory elements detected with epigenetic data	14.29	11.87	8.70	3.73	3.10	2.27
	Exon-associated elements	0.62	0.17	7.38	0.10	0.03	1.13
	Intronic regions	24.00	26.88	19.00	0.90	1.01	0.72
	Intergenic regions	11.17	12.70	24.41	0.21	0.24	0.47
FAN2	Coding variants	11.07	14.16	1.25	29.83	38.16	3.37
	eQTLs detected in muscle	1.84	4.75	0.02	43.58	112.57	0.44
	eQTLs detected in other tissues	13.56	13.38	5.01	69.61	68.65	25.74
	Regulatory elements detected in muscle	14.13	14.68	19.19	19.81	20.59	26.90
	Regulatory elements detected in other tissues	19.09	8.94	4.36	1.40	0.65	0.32
	Exon-associated elements	10.88	13.57	16.76	29.32	36.58	45.17
	Intronic regions	15.81	18.13	1.58	74.71	85.69	7.46
	Intergenic regions	21.32	16.68	9.27	2.81	2.20	1.22

Table S4. Estimated %SNP heritability (proportion of genetic variance explained by a category) and enrichment levels (relative variant effect size per category) for different functional categories with two annotation models, FAN1 and FAN2, when applied to buttock muscling (side view). MGFBLUP models were applied with GRMs constructed with either centered or standardized genotypes, respectively. The annotation categories are described in the Methods section.

	Annotation Group (compartment where variants are located)	%SNP heritability			Enrichment		
		MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC	MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC
FAN1	Coding variants	6.77	4.70	10.71	18.24	12.68	28.87
	eQTLs	13.13	15.28	2.73	62.06	72.25	12.89
	Regulatory elements detected with both techniques	13.11	10.42	0.76	4.42	3.51	0.26
	Regulatory elements detected by ATAC-Seq	24.32	19.43	19.15	3.21	2.56	2.52
	Regulatory elements detected with epigenetic data	5.75	10.35	15.65	1.50	2.70	4.08
	Exon-associated elements	0.02	0.02	6.82	0.00	0.00	1.05
	Intronic regions	18.78	21.89	20.28	0.71	0.82	0.76
	Intergenic regions	18.11	17.90	23.89	0.35	0.34	0.46
FAN2	Coding variants	7.34	5.79	0.01	19.77	15.61	0.03
	eQTLs detected in muscle	2.21	0.03	0.26	52.39	0.80	6.05
	eQTLs detected in other tissues	10.07	12.86	6.34	51.67	65.99	32.56
	Regulatory elements detected in muscle	8.21	8.09	15.78	11.51	11.34	22.13
	Regulatory elements detected in other tissues	32.72	34.55	26.01	2.39	2.53	1.90
	Exon-associated elements	0.02	0.03	11.40	0.00	0.00	1.75
	Intronic regions	19.49	20.95	16.36	0.73	0.79	0.62
	Intergenic regions	19.94	17.70	23.85	0.38	0.34	0.46

Table S5. Estimated %SNP heritability (proportion of genetic variance explained by a category) and enrichment levels (relative variant effect size per category) for different functional categories with two annotation models, FAN1 and FAN2, when applied to buttock muscling (rear view). MGFBLUP models were applied with GRMs constructed with either centered or standardized genotypes, respectively. The annotation categories are described in the Methods section.

	Annotation Group (compartment where variants are located)	%SNP heritability			Enrichment		
		MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC	MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC
FAN1	Coding variants	1.86	2.80	11.84	5.02	7.54	31.91
	eQTLs	15.32	19.30	3.05	72.43	91.23	14.43
	Regulatory elements detected with both techniques	2.06	0.00	14.40	0.70	0.00	4.85
	Regulatory elements detected by ATAC-Seq	14.09	2.60	13.69	1.86	0.34	1.80
	Regulatory elements detected with epigenetic data	10.97	12.17	10.72	2.86	3.17	2.80
	Exon-associated elements	0.32	0.08	0.01	0.05	0.01	0.00
	Intronic regions	38.87	43.94	24.29	1.46	1.65	0.91
	Intergenic regions	16.50	19.11	21.99	0.32	0.37	0.42
FAN2	Coding variants	1.67	3.15	0.00	4.51	8.48	0.00
	eQTLs detected in muscle	6.08	4.82	0.00	144.01	114.10	0.06
	eQTLs detected in other tissues	9.86	14.61	3.27	50.60	75.01	16.77
	Regulatory elements detected in muscle	4.77	1.33	12.38	6.69	1.86	17.35
	Regulatory elements detected in other tissues	21.75	16.83	27.21	1.59	1.23	1.99
	Exon-associated elements	0.31	0.16	4.11	0.05	0.03	0.63
	Intronic regions	38.39	41.86	28.25	1.45	1.58	1.06
	Intergenic regions	17.17	17.24	24.78	0.33	0.33	0.48

Table S6. Estimated %SNP heritability (proportion of genetic variance explained by a category) and enrichment levels (relative variant effect size per category) for different functional categories with two annotation models, FAN1 and FAN2, when applied to global muscling score. MGFBLUP models were applied with GRMs constructed with either centered or standardized genotypes, respectively. The annotation categories are described in the Methods section.

	Annotation Group (compartment where variants are located)	%SNP heritability			Enrichment		
		MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC	MGFBLUP Centered	MGFBLUP Standardized	BayesRR-RC
FAN1	Coding variants	8.01	12.11	0.63	21.59	32.62	1.70
	eQTLs	12.78	13.88	8.65	60.43	65.63	40.91
	Regulatory elements detected with both techniques	21.45	23.48	15.71	7.23	7.91	5.29
	Regulatory elements detected by ATAC-Seq	19.47	8.91	13.21	2.57	1.17	1.74
	Regulatory elements detected with epigenetic data	1.29	0.04	10.83	0.34	0.01	2.82
	Exon-associated elements	0.06	0.03	12.14	0.01	0.01	1.87
	Intronic regions	24.15	28.36	20.29	0.91	1.07	0.76
	Intergenic regions	12.78	13.18	18.54	0.25	0.25	0.36
FAN2	Coding variants	8.02	13.08	15.97	21.62	35.23	43.03
	eQTLs detected in muscle	5.64	5.93	0.24	133.68	140.44	5.64
	eQTLs detected in other tissues	9.46	8.60	3.76	48.57	44.16	19.29
	Regulatory elements detected in muscle	11.00	12.25	6.94	15.42	17.18	9.73
	Regulatory elements detected in other tissues	29.79	23.27	20.64	2.18	1.70	1.51
	Exon-associated elements	0.06	0.04	12.38	0.01	0.01	1.90
	Intronic regions	22.52	24.93	19.72	0.85	0.94	0.74
	Intergenic regions	13.50	11.90	20.36	0.26	0.23	0.39

Table S7. Gain of reliability achieved with BSLMM when using panel FUN1 and LD99. Reliabilities were compared to those obtained with GBLUP using centered genotype and medium density array. Significance levels were estimated by 10,000 bootstraps.

Traits	FUN1		LD99	
	Gain of reliability	P-value	Gain of reliability	P-value
Top	0.052	<0.0001	0.043	<0.0001
Shoulder	0.038	0.0026	0.031	0.0082
Buttock (rear)	0.038	0.0014	0.033	0.0048
Buttock (side)	0.036	<0.0001	0.029	0.0008
Global score	0.049	<0.0001	0.043	<0.0001

Table S8. Reliability obtained using whole-genome sequence data, with or without annotation, compared to that obtained using a medium marker density panel

Panel	Model	Method	Shoulder	Top	Buttock side	Buttock rear	Global score
MMD	Without annotation	GBLUP-C	0.792	0.674	0.686	0.75	0.705
		GBLUP-S	0.785	0.67	0.684	0.741	0.697
		BayesR	0.793	0.678	0.69	0.75	0.707
WGS	Without annotation	GBLUP-C	0.811	0.694	0.7	0.758	0.724
		GBLUP-S	0.799	0.681	0.697	0.749	0.712
		BayesR	0.806	0.695	0.71	0.766	0.726
WGS	FAN1	MGFBLUP-C	0.818	0.708	0.718	0.77	0.732
		MGFBLUP-S	0.807	0.695	0.717	0.762	0.72
		BayesRR-RC	0.795	0.69	0.71	0.75	0.715
WGS	FAN2	MGFBLUP-C	0.819	0.708	0.718	0.769	0.734
		MGFBLUP-S	0.807	0.697	0.717	0.761	0.722
		BayesRR-RC	0.794	0.694	0.715	0.753	0.715
WGS	LDMS	MGFBLUP-C	0.802	0.689	0.701	0.748	0.717
		MGFBLUP-S	0.799	0.688	0.702	0.749	0.719
		BayesRR-RC	0.774	0.67	0.686	0.729	0.689
WGS	LDMS x FAN1	BayesRR-RC	0.758	0.666	0.683	0.709	0.677

Table S9. Reliability of different methods when using different marker panels selected via functional annotation or with an LD pruning strategy

Traits	Method	WGS	LD99	LD98	LD95	LD85	LD80	FUN1	FUN2	FUN3	ARRAY
Shoulder	BayesCπ	NA	0.795	0.795	0.792	0.785	0.782	0.795	0.799	0.784	0.809
	BayesR	0.806	0.806	0.804	0.803	0.795	0.792	0.807	0.809	0.793	0.815
	BSLMM	NA	0.823	0.822	0.82	0.811	0.814	0.83	0.83	0.818	0.82
	GBLUP-C	0.811	0.809	0.808	0.806	0.804	0.803	0.811	0.813	0.8	0.815
	GBLUP-S	0.799	0.796	0.794	0.792	0.788	0.786	0.796	0.799	0.781	0.811
Top	BayesCπ	NA	0.685	0.68	0.677	0.676	0.674	0.685	0.684	0.668	0.691
	BayesR	0.695	0.697	0.695	0.695	0.686	0.685	0.701	0.7	0.691	0.701
	BSLMM	NA	0.718	0.717	0.716	0.712	0.712	0.726	0.724	0.716	0.711
	GBLUP-C	0.694	0.695	0.694	0.693	0.691	0.691	0.701	0.699	0.689	0.697
	GBLUP-S	0.681	0.683	0.681	0.679	0.675	0.673	0.684	0.684	0.67	0.692
Buttock side	BayesCπ	NA	0.703	0.703	0.705	0.7	0.7	0.715	0.714	0.716	0.707
	BayesR	0.71	0.716	0.716	0.715	0.713	0.712	0.721	0.72	0.723	0.715
	BSLMM	NA	0.716	0.716	0.716	0.715	0.715	0.722	0.72	0.724	0.711
	GBLUP-C	0.7	0.707	0.708	0.707	0.706	0.706	0.716	0.715	0.714	0.703
	GBLUP-S	0.697	0.704	0.704	0.703	0.701	0.699	0.714	0.711	0.714	0.704
Buttock rear	BayesCπ	NA	0.755	0.749	0.752	0.744	0.752	0.751	0.752	0.746	0.766
	BayesR	0.766	0.767	0.766	0.765	0.759	0.758	0.768	0.767	0.763	0.773
	BSLMM	NA	0.783	0.78	0.782	0.78	0.778	0.787	0.786	0.784	0.778
	GBLUP-C	0.758	0.764	0.764	0.764	0.763	0.762	0.764	0.765	0.758	0.767
	GBLUP-S	0.749	0.753	0.753	0.752	0.75	0.747	0.753	0.752	0.746	0.765
Global score	BayesCπ	NA	0.711	0.709	0.702	0.71	0.708	0.711	0.712	0.705	0.727
	BayesR	0.726	0.729	0.727	0.727	0.723	0.72	0.728	0.728	0.723	0.735
	BSLMM	NA	0.748	0.746	0.743	0.74	0.739	0.754	0.754	0.744	0.746
	GBLUP-C	0.724	0.725	0.725	0.724	0.723	0.722	0.725	0.726	0.716	0.731
	GBLUP-S	0.712	0.713	0.712	0.711	0.708	0.706	0.712	0.712	0.703	0.726

Table S10. Reliability obtained using marker panels selected on the basis of functional information and for models with or without the incorporation of functional annotation

Panel	Approach	Model	Shoulder	Top	Buttock side	Buttock rear	Global score
FUN1	GBLUP-C	Without annotation	0.811	0.701	0.716	0.764	0.725
	BayesR	Without annotation	0.807	0.701	0.721	0.768	0.728
	MGFBLUP-C	FAN1	0.812	0.701	0.713	0.766	0.726
	BayesRR-RC	FAN1	0.795	0.691	0.711	0.755	0.712
	MGFBLUP-C	FAN2	0.807	0.699	0.713	0.764	0.726
	BayesRR-RC	FAN2	0.793	0.692	0.714	0.76	0.714
FUN2	GBLUP-C	Without annotation	0.813	0.699	0.715	0.765	0.726
	BayesR	Without annotation	0.809	0.7	0.72	0.767	0.728
	MGFBLUP-C	FAN1	0.812	0.701	0.714	0.766	0.726
	BayesRR-RC	FAN1	0.798	0.692	0.711	0.759	0.715
	MGFBLUP-C	FAN2	0.806	0.696	0.714	0.764	0.726
	BayesRR-RC	FAN2	0.792	0.689	0.714	0.759	0.715
FUN3	GBLUP-C	Without annotation	0.8	0.689	0.714	0.758	0.716
	BayesR	Without annotation	0.793	0.691	0.723	0.763	0.723
	MGFBLUP-C	FAN1	0.807	0.698	0.709	0.761	0.722
	BayesRR-RC	FAN1	0.794	0.692	0.716	0.755	0.713
	MGFBLUP-C	FAN2	0.806	0.698	0.708	0.761	0.722
	BayesRR-RC	FAN2	0.793	0.69	0.714	0.756	0.71

5.9 Supplementary figures

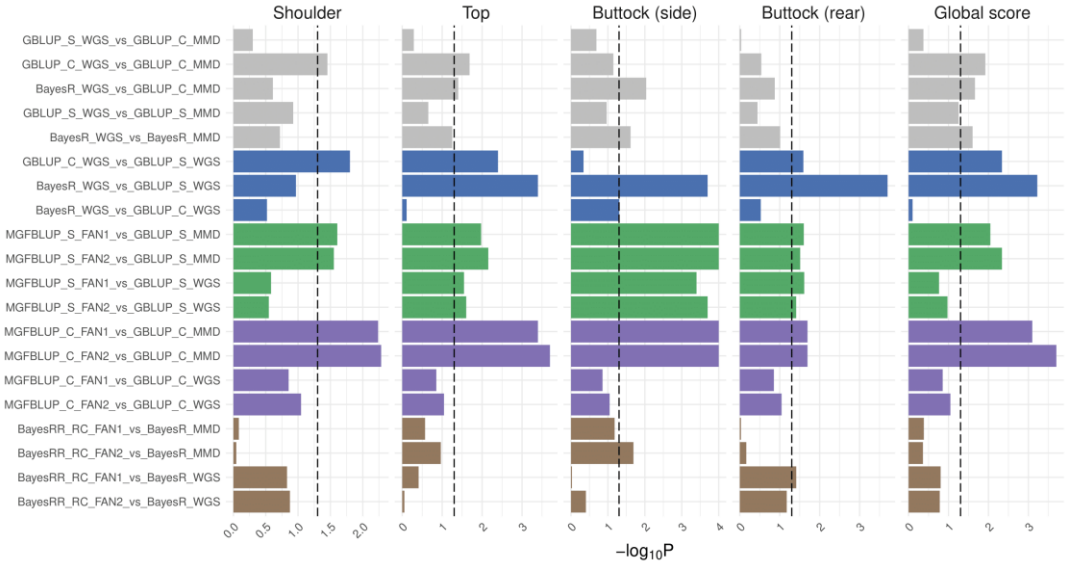


Figure S1. Significance levels of difference in reliabilities obtained with methods using whole-genome sequence data, with or without annotation. Comparisons were also made with medium marker density (MMD) array prediction. The names of the pair of methods compared are indicated on the left, with the extensions '-C' and '-S' indicating whether the GRMs used in GBLUP and MGFBLUP models were constructed with centered and standardized genotypes, respectively, WGS referring to the use of the whole-genome sequence data without annotation, FAN1 and FAN2 referring to the two models incorporating functional annotation (describing in Table 1). P-values of differences were obtained by bootstrapping and are presented on a -log10 scale, the dashed line indicates the significance threshold at p=0.05. The colors facilitate the reading of the results and indicate which pairs of methods are compared: gray for comparisons between WGS and MMD, blue for comparisons of the three methods at the sequence level, green for comparisons of GBLUP-S (without functional annotation) and MGFBLUP-S (with functional annotation) models, purple and brown for the same comparisons with GBLUP-C versus MGFBLUP-C models and with BayesR versus BayesRR-RC models, respectively.

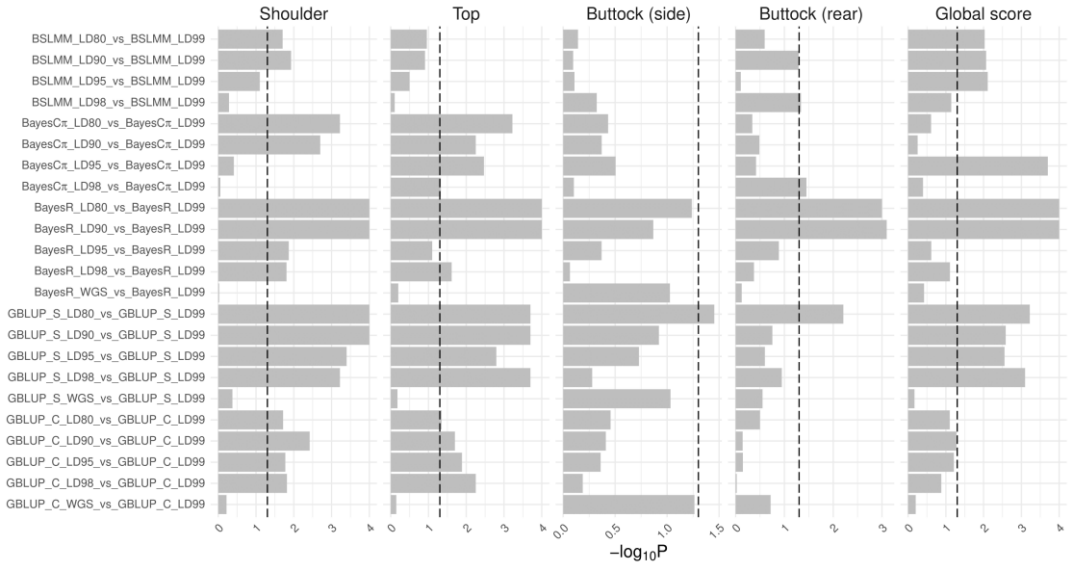


Figure S2. Significance levels of the difference in reliabilities obtained for each method when using subsets of markers selected based on LD pruning with different thresholds. The comparisons are made per method, with different selected subsets of markers. The names of the compared method pairs are given on the left, where GBLUP-S and GBLUP-C refer to GBLUP with centered and standardized genotypes, WGS refers to the use of whole-genome sequence data, and LD80 to LD99 refer to marker panels selected by LD pruning with the threshold set at $r^2 > 80$ to 99, respectively. P-values of differences were obtained by bootstrapping and are presented on a -log10 scale, the dashed line indicates the significance threshold at p=0.05.

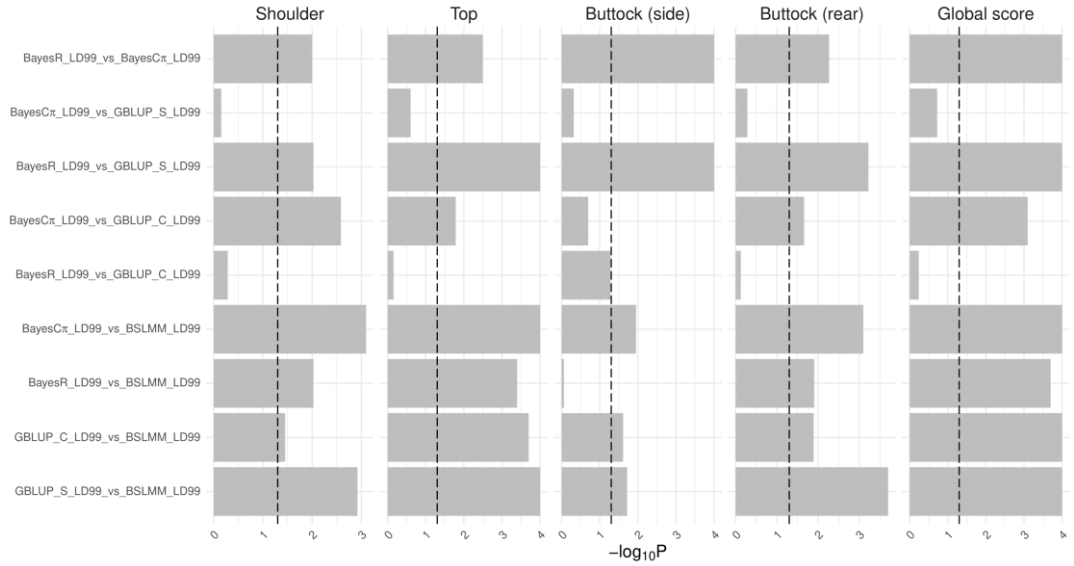


Figure S3. Significance levels of the difference in reliabilities obtained between different methods when using a subset of markers selected based on LD pruning with of threshold at $r^2 > 0.99$. The names of the compared method pairs are given on the left, where GBLUP-S and GBLUP-C refer to GBLUP with centered and standardized genotypes, LD99 refers to marker panels selected by LD pruning with the threshold set at $r^2 > 99$. P-values of differences were obtained by bootstrapping and are presented on a -log10 scale, the dashed line indicates the significance threshold at p=0.05.

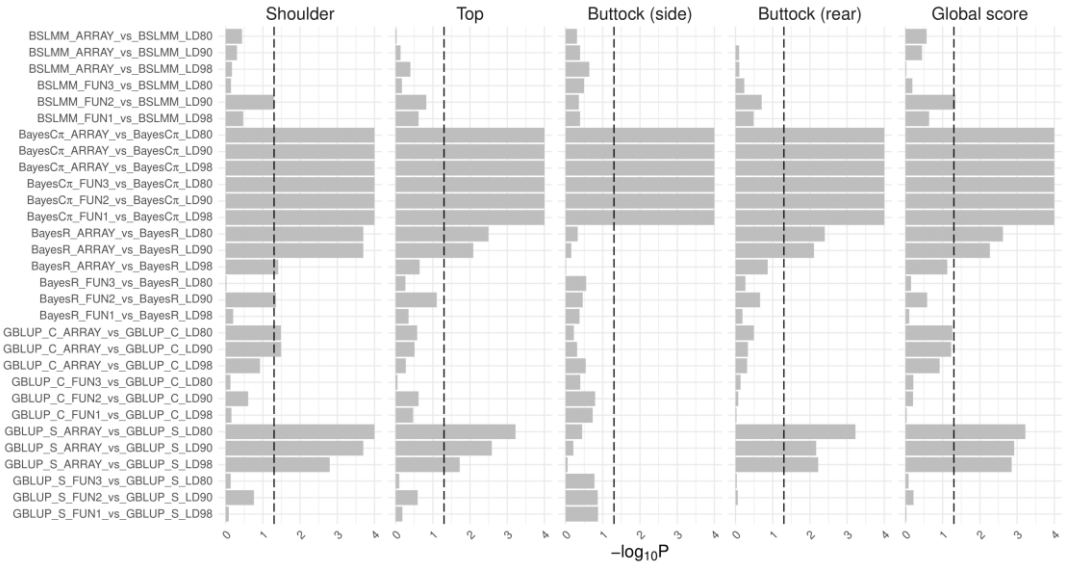


Figure S4. Significance levels of the difference in reliability obtained for each method when using subsets of markers selected based on functional annotation, LD pruning or their presence on commercial arrays. The comparisons are made per method, with different selected subsets of markers. The names of the compared method pairs are given on the left, where the extensions '-C' and '-S' indicate whether the GRM used in that GBLUP model was constructed with centered and standardized genotypes, respectively, FUN1 to FUN3 refer to marker panels selected based on functional annotation, ARRAY refers to markers present on commercial genotyping arrays, and LD80, LD90, and LD98 refer to marker panels selected by LD pruning with the threshold set at r^2 > 80, 90, and 98, respectively (see Table 2 for more details on the marker panels). FUN and LD panels were compared for panels of approximately the same size, with the number of markers on the ARRAY panel being approximately the same as on the FUN3 and LD80 panels. P-values of differences were obtained by bootstrapping and are presented on a -log10 scale, the dashed line indicates the significance threshold at p=0.05.

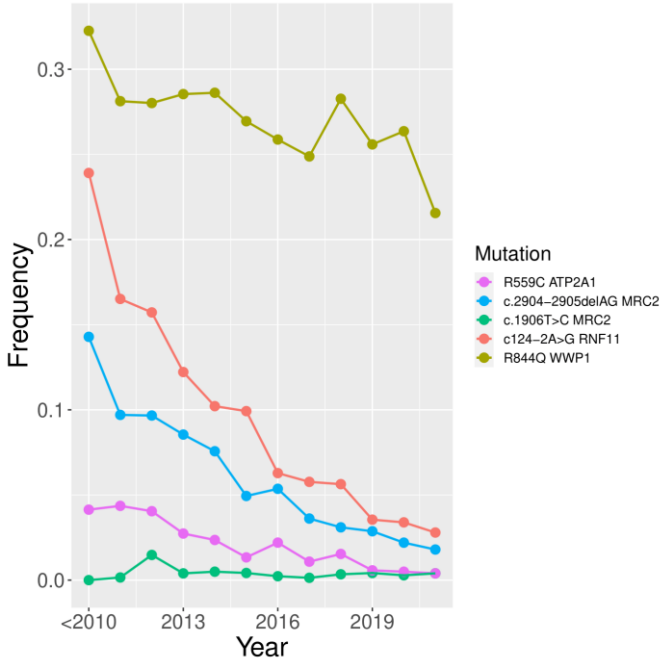


Figure S5. Evolution of the frequency of five known recessive deleterious variants with positive effect on muscular traits in heterozygotes. These five recessive deleterious variants associated with congenital muscular dystonia (CMD), crooked tail syndrome (CTS) or stunted growth have previously been shown to confer a heterozygote advantage. Heterozygotes do indeed have higher muscular development. The reduction in frequency suggests that although these large effect variants account for a substantial proportion of the genetic variance in the reference population, their influence in the target population is now much less.

Discussion - Perspectives

6 Discussion - Perspectives

6.1 Identifying the Bovine Regulatory Element and Perspectives

Extensive research in humans has highlighted the important role of regulatory variants in shaping complex traits (Boix et al., 2021; Meuleman et al., 2020; Trynka et al., 2015). However, the detection of regulatory elements in cattle has primarily been limited to a few tissues and developmental stages, preventing a comprehensive understanding of the full spectrum of regulatory elements and their activity (Foissac et al., 2019; Halstead et al., 2020a, 2020b; Kern et al., 2021; Ming et al., 2021; Powell et al., 2023). During my PhD, I contributed to the generation of a comprehensive catalog of regulatory elements using 104 samples covering 63 bovine tissue types. For tissues with a major role in agronomically important traits, such as the mammary gland, we included multiple individual samples to better capture their diversity. To my knowledge, this is the most comprehensive catalog of open chromatin regions in cattle to date, including novel tissues such as the immune and digestive systems. The entire catalogue accounts for approximately 10% of genome, in line with findings in humans. We defined core and consensus regions of open chromatin following Meuleman et al. (2020). Core regions are detected as open in more tissues, ensuring the authenticity of the signal and indicating activation across multiple tissue types. They also showed greater enrichment of eQTL signals compared to the broader consensus regions, consistent with Meuleman's observation of higher enrichment of genetic signals associated with complex traits in their core regions. To build a comprehensive map of regulatory elements, we also integrated results from previous studies by downloading publicly available data. We demonstrated the benefit of adding additional tissues to discover new peaks and obtained a saturated map when more than 90 tissue samples were included. Although most of the data from other studies did not pass our quality checks, some tissues, such as embryonic tissue from Halstead et al. (2020b), provided substantial information. Indeed, embryonic tissues showed a markedly different open chromatin landscape compared to other tissues, with approximately 200,000 specific peaks that were absent in other tissue types. This highlights the importance of including different tissue types and developmental stages to comprehensively mine regulatory elements. Based on the genomic locations of ATAC-Seq peaks, we classify them into distal and proximal regulatory elements, which are characterized by different roles and mechanisms. Previous studies have shown that proximal regulatory elements are mainly promoters, while distal elements include enhancers, insulators, and silencers, among others (Maston et al., 2006). In the present work, we found that proximal peaks are much less abundant than distal peaks. Although a possible explanation is the smaller genomic space used to define proximal peaks, this alone cannot explain the difference. Other factors, such as the greater abundance of enhancers compared to promoters, as highlighted by previous ENCODE studies, may also play a role (Field and Adelman, 2020). Compared to distal peaks, proximal peaks are larger and more accessible. Interestingly, proximal peaks are active in more tissues than distal peaks, suggesting that distal peaks are more responsible for tissue-specific activity. This is consistent with previous studies indicating that

distal regulatory elements exhibit high tissue specificity (Liu et al., 2017). In contrast to previous studies, the inclusion of a wide range of tissue types covering different biological processes allowed us to interpret the biological activity of regulatory elements. Using the method proposed by Meuleman et al (2020), we compressed the matrix indicating the presence of each peak in all samples into 16 main factors representing different biological activities. We then inferred peak functions based on their dominant components (i.e. tissues). This information could, for example, help to identify causative regulatory variants associated with specific complex traits associated with these tissues. Among these dominant components, 12 were clearly associated with tissues that could be readily assigned to recognizable body systems corresponding to different biological processes, accounting for a total of 64.5% of the peaks. For example, we identified 58,078 peaks assigned to the mammary gland-associated component, suggesting a role in lactation, and 49,988 muscle-related peaks that are likely to be involved in muscle growth and development. Figure 6.1 shows two examples of how such information is relevant to the identification of tissue-specific regulatory variants. These correspond to two open chromatin segments close to the lactalbumin alpha gene (*LALBA*) and myosin heavy chain 1 (*MYH1*) genes, which have been assigned to the mammary-gland and muscle components, respectively. The availability of different tissue types has the advantage of clearly demonstrating that these signals are specifically detected in the corresponding tissues.

Our catalog and its annotation are a valuable resource for the scientific community. Our link, available on the UCSC Genome Browser, facilitates the visualization of regulatory elements and their function, providing insight into their implications in genetics, evolution, and related fields. In our dataset, most regulatory elements exhibited highly dynamic activity across tissues, with only a small fraction of peaks being ubiquitous, suggesting that specific peaks may play critical roles in tissue-specific functions. This again highlights the importance of including a variety of tissue types to adequately capture these functionally specific peaks. Tissue-specific regulation has also been observed in other breeds, such as indicine cattle (Alexandre et al., 2021), while it has also been reported that these tissue-specific regulations are conserved across vertebrates (Kern et al., 2021). Our results show that over 213,000 peaks are specific to embryonic stages and their accessibility has been reported to change dramatically during development, for example between 2-4 cell embryos and morula stage embryos (Halstead et al., 2020b; Ming et al., 2021). This highlights the critical role of regulatory activity in development and differentiation, particularly in early life stages. Importantly, changes in chromatin accessibility also play a critical role in regulating gene expression in response to stress and environmental changes (Boschiero et al., 2022; Fang et al., 2019; Johnston et al., 2021). Therefore, future research efforts will need to include samples across all developmental stages and environmental conditions to effectively elucidate stage- and context-specific gene regulation, requiring dynamic data sampling strategies. In addition, recent studies comparing regulatory elements associated with the bovine immune system across breeds have highlighted the benefits of including individuals with diverse genetic backgrounds (Powell et al., 2023). Besides adding more samples as described above, future

studies may benefit from complementary techniques. While ATAC-Seq stands out as the most efficient method for uncovering regulatory elements, its combination with ChIP-seq of other epigenetic marks could significantly improve our understanding of their specific regulatory roles by inferring chromatin states (Ernst and Kellis, 2012). Furthermore, these combined approaches may help to identify regulatory elements that ATAC-Seq alone may miss, such as the unbound site of CTCF (Oomen et al., 2019). For example, Kern et al. (2021) combined ATAC-Seq data with other histone marks to identify the transcription factor footprints within regulatory elements, showing significant enrichment of known motifs. Furthermore, the integration of ATAC-Seq data from larger numbers of individuals with other omics datasets, such as transcriptomics, would allow the study of the association of gene expression with open chromatin information, thereby facilitating the detection of causal variants perturbing gene expression and the deciphering of gene regulation (Boix et al., 2021; Kim et al., 2024; Kumasaka et al., 2016). The inclusion of a large number of individuals with different phenotypes and for multiple tissues, together with other omics data, will be crucial to link phenotypes in different environments to regulatory elements that control the gene response to stress (Alasoo et al., 2018; Arthur et al., 2024; Boix et al., 2021; Kumasaka et al., 2016; Meuleman et al., 2020). However, it seems impossible to perform this labor- and cost-intensive task in a single laboratory. It is clear that a collaborative effort is required to generate the large-scale data needed to systematically identify regulatory elements and investigate their dynamic changes across tissues, developmental stages, individuals and environments. For these reasons, several major consortia such as BovReg and FAANG have been established (Andersson et al., 2015; Moreira et al., 2022). These have already generated large datasets covering more tissues from more individuals and in more conditions, as well as additional epigenetic data, and their first results show promise for the future. Finally, the recent advent of single-cell technology combined with ATAC-Seq (scATAC-Seq) (Buenrostro et al., 2015; Hu et al., 2023; Mezger et al., 2018) or ChIP-seq (Grosselin et al., 2019; Rotem et al., 2015) approaches provides new opportunities to study cell type-specific regulatory elements that control distinct cellular functions and contribute to cellular diversity. These new approaches have already allowed us to map causal variants to specific cell populations (Lake et al., 2018) and to distinguish the regulatory networks of different cell types in the same microenvironment, such as malignant and immune cells of basal cell carcinoma (Satpathy et al., 2019). In cattle, the availability of scATAC-Seq data is gradually expanding, allowing studies of regulatory elements for specific cell populations, such as cell types involved in processes such as bovine skeletal muscle development (Cai et al., 2023), immune responses (Gao et al., 2022b; Wang et al., 2023), and somatic cell nuclear transfer (Huang et al., 2023). In the future, ongoing advances in single-cell technologies hold the promise that integrating data on transcriptional and epigenetic states at the single-cell level will greatly enhance our understanding of the regulatory dynamics underlying complex traits.

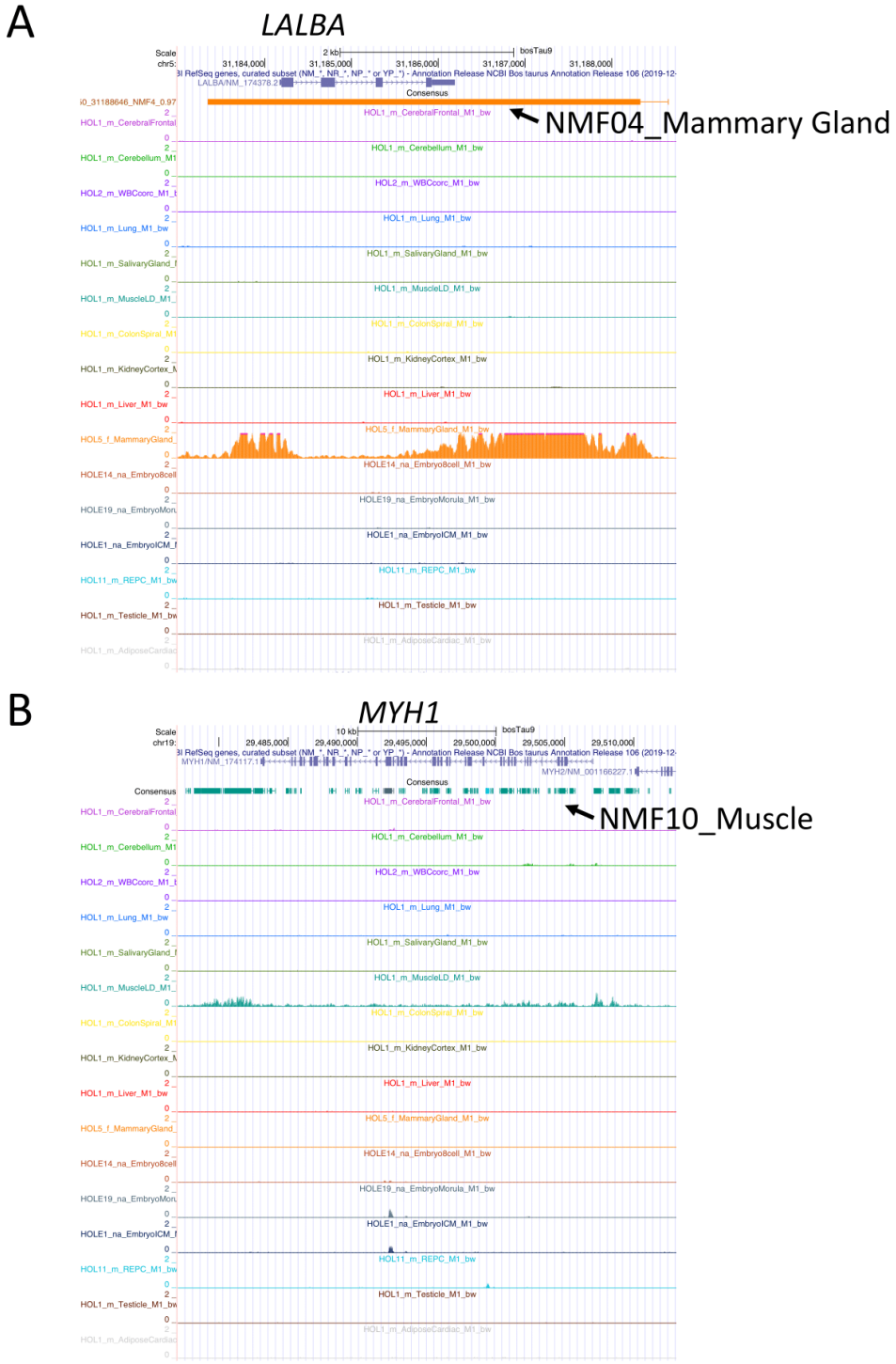


Figure 6.1. UCSC Genome Browser view of tissue-type-specific ATAC-Seq peaks and their Negative Matrix Factor based annotations (https://www.gigauag.uliege.be/cms/c_4791343/en/gigauag-diagnostics-software-data) around (A) a mammary gland specific gene, LALBA (Lactalbumin Alpha) and (B) a muscle specific gene, MYH1 (Myosin Heavy Chain 1) where chromatin is opened in a tissue specific way across the genes.

6.2 Characteristics of Variants Located within ATAC-Seq Peaks

To lay the groundwork for using regulatory variants to enhance genomic selection, we characterized variants mapping to ATAC-Seq peaks in 264 Dutch Holstein Friesians. A total of 1,390,391 variants were identified, of which 67% were common variants with a minor allele frequency (MAF) > 0.05. These ATAC-Seq peaks are expected to be enriched for regulatory elements that are functionally important. Therefore, they should have lower levels of polymorphism and greater evolutionary conservation compared to other regions of the genome. We observed an over-dispersed distribution of GERP scores in ATAC-Seq peaks, suggesting a higher proportion of constrained elements, consistent with numerous studies indicating that regulatory regions are evolutionarily conserved (Kern et al., 2021). For example, a recent comparative genomic study of central placental mammals (Christmas et al., 2023) found that most constrained bases in the genome are located within regulatory elements, while coding regions are the most enriched in constrained bases (Sullivan et al., 2023). However, this was accompanied by a higher proportion of less constrained regions, supporting an accelerated mutation rate. Compared to the flanking regions, we indeed observed a higher mutation rate in these regions, which has not been previously reported in cattle or even in livestock. We hypothesized that this was due to a higher mutation rate in ATAC-Seq regions and rigorously tested this using a variety of approaches. A recent study of mutation rates in Arabidopsis revealed epigenomic features that correlate with mutation rates, with open chromatin emerging as one of the most highly correlated features associated with elevated mutation rates (Monroe et al., 2022). This observation has also been made in human cells, such as somatic cells (Luquette et al., 2022) or human spermatogonia (Kaiser et al., 2021). One possible explanation for this is that the reduced efficiency of Pol-δ-mediated displacement of error-prone Pol-α-synthesized primers is due to the presence of a barrier on the DNA, such as a DNA-bound protein (Reijns et al., 2015). During DNA replication, double-stranded DNA is separated into two single strands by the replicative helicase, forming a replication fork (Leman and Noguchi, 2013). New strands are synthesized using these single strands as templates. The new DNA synthesized in the 5’ to 3’ direction is known as the leading strand, which uses the 3’ to 5’ template strand. This synthesis proceeds in the same direction as the replication fork, allowing it to be continuous and requiring only one primer at each replication fork. In contrast, the other strand, called the lagging strand, is synthesized in the opposite direction of the growing replication fork (Snedeker et al., 2017). As a result, multiple primers are used and many small fragments, known as Okazaki fragments, are synthesized in parallel. These fragments are later linked together to form the lagging strand (Figure 6.2). The primers for the leading strand and the Okazaki fragments are synthesized by different DNA polymerases, with Pol-α responsible for the synthesis of primers for the Okazaki fragments (Snedeker et al., 2017). However, Pol-α has low fidelity and a higher error rate due to its lack of 3’ to 5’ proofreading exonuclease activity. During replication (Snedeker et al., 2017), the DNA primers synthesized by Pol-α are removed and replaced by more accurate DNA fragments synthesized by Pol-δ.

Notably, the rapid binding of DNA-binding proteins to the Okazaki fragments can interfere with this replacement process, resulting in the retention of primers synthesized by the low-fidelity Pol-α (Figure 6.3). The retention of the fragment synthesized by Pol-α is thought to lead to a higher mutation rate in the binding sites of the regulatory factors (Reijns et al., 2015). In addition, DNA-binding proteins such as transcription factors have been reported to interfere with nucleotide excision repair, which may be another reason for the increased mutation rate in the regulatory regions (Sabarinathan et al., 2016). Our data do not support the hypothesis that the number of variants in DNA binding sites identified by their associated motifs is higher. We hope that in the future, more motifs of DNA binding sites will be obtained from experimental approaches, providing more accurate results compared to the in silico predicted motif binding sites we used. Overall, the observed higher mutation rate remains inconsistent with the functional importance of regulatory regions. To further investigate this apparent contradiction, we compared the site frequency spectrum (SFS) of different categories of variants and found that variants in regulatory regions are under purifying selection, supporting their functional importance. A higher selection strength was observed for indels compared to SNVs. Variants in regulatory elements thus appear to be under more complex selection pressures and mutation rates than we initially thought. This suggests that conservation scores are only useful to identify a fraction of regulatory elements that are highly conserved across species and under strong selection. Experimental methods such as ATAC-Seq will therefore remain essential to identify species-specific regulatory elements or those that are less conserved as a result of weak selection. These elements may exhibit higher sequence diversity across species due to a higher mutation rate.

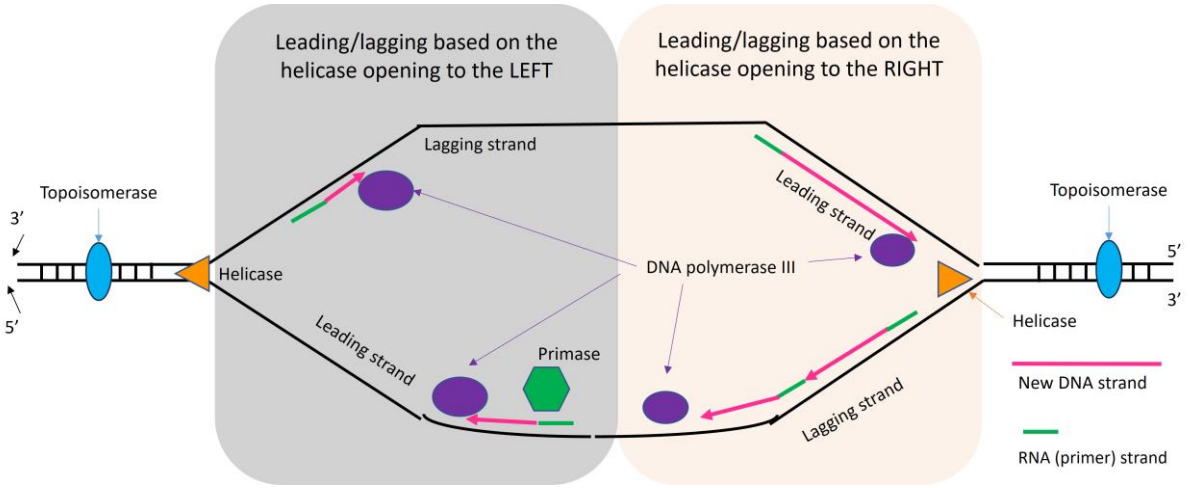


Figure 6.2. Process of DNA replication. Double-strand DNA first is unwound by helicase (orange triangle) from the center to both sides, forming a bubble known as the replication fork. Then the leading strand and the Okazaki fragment are synthesized using the primer shown in green. For the lagging strand, the primer is synthesized by DNA polymerases, Pol-α. This image has been copied from https://passel2.unl.edu/view/lesson/6f214d098527/13

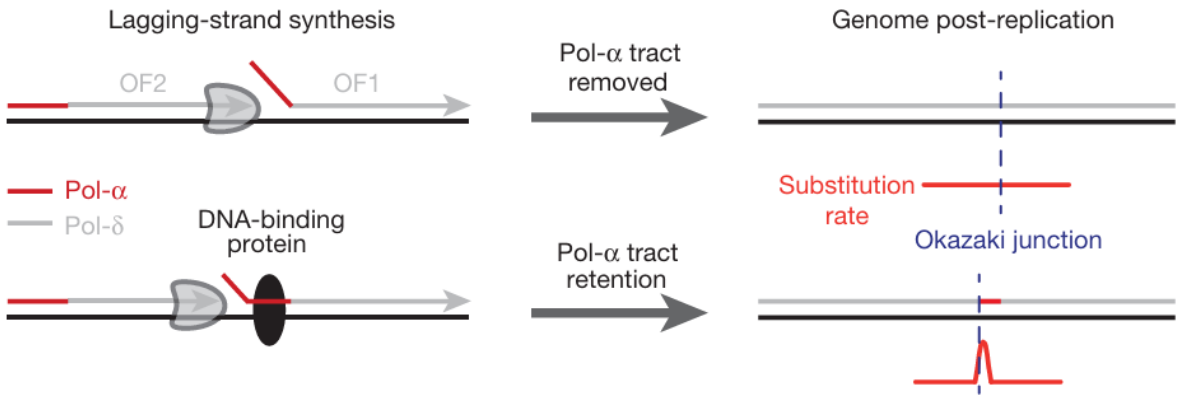


Figure 6.3. Schematic representation of the Pol-α tract retention hypothesis. During replication, the lagging strand, DNA at the beginning of the Okazaki fragment, is synthesized by the low-fidelity polymerase Pol-α (the red line in the left panel). It is replaced by the new DNA fragment synthesized by Pol-δ when linking the Okazaki fragment shown in the upper right panel, but remain in the lagging strand when this primer is occupied by DNA-binding protein that increase the mutiont rate compared to the flank region. This figure has been obtained from Reijns et al. (2015).

6.3 Enrichment of regulatory variants in ATAC-Seq peaks

To estimate the enrichment of regulatory variants in ATAC-Seq peaks, we first conducted eQTL analyses in two tissues, blood and liver, to detect regulatory variants. In total, we identified 7817 and 6172 eQTLs and their credible sets in blood and liver, respectively. These credible sets are assumed to include causative variants as well as variants in LD with causative variants. Using Trynka's method (Trynka et al., 2015), we demonstrated that these credible sets are indeed more frequent in ATAC-Seq peaks. Importantly, we showed that this enrichment is not due to positional confounding, such as proximity to transcription start sites. Of particular interest, the overlap with the tissue-specific NMF was predominantly associated with distant regulatory elements. This finding is consistent with previous studies indicating that the distal regulation of gene expression has pronounced tissue-specific features (Liu et al., 2017). However, there is no difference between distant and proximal regulatory elements in terms of overlap with the ubiquitous NMF. To further estimate the proportion of regulatory variants mapped to regulatory regions and the proportion of variants within regulatory regions that are truly regulatory, we developed a maximum likelihood approach integrating eQTL and ATAC-Seq data. We estimated that only one third of the regulatory variants mapped in the existing ATAC-Seq catalog, which already covers almost all tissue types. A plausible explanation is the mismatch in developmental stages between our ATAC-Seq catalog and the eQTL data. Therefore, future studies including samples from different developmental stages are essential to elucidate the dynamics of regulatory elements during development. In addition, other factors may explain our observation. For example, ATAC-Seq may not capture all regulatory elements, potentially missing small regions that are open but not effectively amplified, silencers/insulators that are sometimes undetectable, or eQTL signals associated with splicing mutations. The inefficiency of ATAC-Seq in silencer detection was illustrated by the lack of ATAC-

Seq signal associated with the regulatory variant previously identified in *IGF2*. Although 24 out of 25 variants in ATAC-Seq are non-regulatory, this knowledge is still useful for identifying causative regulatory variants. It is also possible that incorporating eQTL data from other tissues would increase the proportion of variants in ATAC-Seq estimated to be regulatory, but publicly available eQTL resources such as cGTEx rely on genotype imputation based on RNA-seq data, which may not meet the requirements of our study.

6.4 Evaluation of heritability partitioning methods in livestock

Heritability partitioning plays a central role in investigating the genetic architecture underlying complex traits, allowing a deeper understanding of the genetic factors associated with these traits, and providing opportunities to improve genomic selection (Loh et al., 2015a; Yang et al., 2011b). Currently, heritability partitioning methods have been mainly developed and validated using human data (Finucane et al., 2015; Patxot et al., 2021; Yang et al., 2015), but they are widely used in livestock genetic studies with limited validation. Livestock species, such as cattle, are subject to intensive selection and exhibit significant differences in demographic history and genomic characteristics compared to humans. These differences include small effective population size (Hayes et al., 2003; MacLeod et al., 2013), high levels of inbreeding (Leroy, 2014) and relatedness, and long-range linkage disequilibrium (Gautier et al., 2007). To date, the GREML approach has only been evaluated for heritability partitioning in cattle in a single study using less than 3000 individuals and a limited number of scenarios (Cai et al., 2022). Therefore, our study, which evaluates state-of-the-art approaches with a large cohort and using (imputed) sequence data, provides valuable insights into the properties of heritability partitioning methods and their accuracy in different scenarios, both with and without stratification, when applied to livestock. Overall, our results indicate lower levels of precision in cattle compared to human studies, for both GREML and BayesRR-RC methods. Higher standard errors were observed even in the simplest scenarios, where a lower bias was observed for the LDMS model, although all the models we tested provided good estimation. We speculate that a major factor in this discrepancy is the high level of LD and individual relatedness, where differences between individuals are small. In human studies, one individual from each pair with a relatedness greater than 0.025 is typically excluded to ensure higher resolution (Speed et al., 2012). However, using such a threshold in cattle would result in the exclusion of most individuals due to the reduced effective population size resulting from domestication and artificial selection, resulting in higher relatedness between individuals. In scenarios where causative variants were enriched in certain LD score and MAF categories, models accounting for stratification provided unbiased results in agreement with previous studies (Yang et al., 2015), but the methods remained imprecise. This high level of variation in estimates was also observed and even more pronounced in Cai's simulation study in Holstein cattle (Cai et al., 2022), across scenarios with and without MAF stratification, and was particularly evident in the simplest scenarios where we did not use

functional annotation. In addition, we found that for GREML, assuming the same relationship between effect size and LD score or allele frequency to construct the GRM and to simulate the phenotype led to improved results, although the true relationship on real data remains unknown. Interestingly, LDMS models appeared robust when the rules used to construct the GRM were different from those used in the simulations. When enrichment levels of multiple functional categories were estimated simultaneously, accurate heritability partitioning was achieved only in scenarios with no enrichment (equal level for all categories) and with 100% enrichment (a single category contributing to genetic variation). In the most complex scenarios, where multiple functional groups had different enrichment levels, high variation in estimates was observed, especially for small annotation groups. Despite the poor estimation in these complex scenarios, the results remained informative and reflected the true ranking. Strong confounding between certain functional groups, such as variants upstream and downstream of genes and variants in OCR, led to systematic biases in their estimators for both BayesRR-RC and GREML methods. In human studies, this confounding effect has also been observed for annotation groups that are short in length and in high LD with other groups (Gusev et al., 2014). Correlations between elements from the GRMs were used in our study to quantify the degree of confounding between categories and showed that the category of intergenic variants had less similarity to other annotation groups. This may explain why higher accuracies were observed when %SNP heritability was estimated for this category using the two-component model. Other parameters, such as the number of causative variants, SNP heritability, and the distribution of effect sizes in different categories, may also affect the accuracy of parameter estimation. However, it remains difficult to comprehensively simulate all parameter combinations. Nevertheless, compared to previous studies in cattle, our research provides new insights by using finer annotation levels, e.g. by including a large catalog of experimentally obtained functional data, by including more individuals with sequence-level data, and by exploring more complex scenarios.

Several methods have been developed to estimate the levels of enrichment of different functional groups. GREML and BayesRR-RC are the most commonly used methods that use individual data to directly estimate the heritability explained by each category. GREML has been widely used in livestock studies due to its lower computational requirements and its ability to handle genotype data at the sequence level. Recently, methods based on summary statistics have emerged as powerful tools for partitioning heritability in human studies, but are rarely used in livestock (Wray et al., 2019). Recently, lower accuracy was reported in a study conducted in cattle (Xiang et al., 2023). Obtaining summary statistics in livestock species is also computationally demanding, as GWAS must be performed using linear mixed models due to the high levels of relatedness and stratification. BayesRR-RC is the first software to allow partitioning of heritability in large cohorts with sequence-level data using a Bayesian model in a reasonable time (Patxot et al., 2021). Running these two approaches in livestock with different models (e.g., LDMS, MS, LDS, annotation groups) allows understanding whether there is stratification among variant effects and which categories have a higher contribution to the genetic variance of a complex trait. In our study, GREML had a lower standard deviation than BayesRR-RC in

simple scenarios, but BayesRR-RC provided a better estimate with higher resolution in complex scenarios. In particular, the rules used to simulate the phenotypes (i.e., the relationship between effect size and MAF) in the simpler scenarios matched the rule used to construct the GRM, which may have favored GREML. In fact, the performance of GREML began to degrade when using a GRM constructed based on different rules. Unlike LDMS, which corrects for stratification by grouping variants, LDAK takes a quantitative approach by weighting variants based on their LD (Speed et al., 2012). However, the weighting parameters are estimated based on human data and require further study to extrapolate to other species due to different LD patterns in non-human genomes. In our study, when using the LDMS approach with GREML, nearly half of our simulations failed to converge due to the excessive number of parameters to be estimated. This problem has been reported elsewhere (Finucane et al., 2015; Speed et al., 2017), and fitting a large number of categories can also lead to a huge demand on computational resources. While EM-GREML has been reported to converge where AI-GREML fails (Misztal, 2008), its computational inefficiency hinders widespread implementation. In contrast, BayesRR-RC allows the estimation of more hyperparameters. However, when fitting models with fewer categories, GREML shows greater computational efficiency than BayesRR-RC, as sampling in BayesRR-RC becomes more time consuming. In our simulations, the two-component model corresponding to the GFBLUP, which is widely used to include annotations in variance component estimation or genomic prediction models (Edwards et al., 2016), consistently produced biased results in the most complex scenarios, regardless of LDMS correction. Therefore, we advocate the use of more categories instead of just two to mitigate overestimation due to absorption of effects from neighboring categories, especially for categories close to functional categories enriched in causal variants, such as coding sequence. Following the approach of Orliac et al. (2022), we ran 5000 iterations for BayesRR-RC in our analysis, which proved sufficient for parameter convergence, as evidenced by convergence diagnostic plots showing detailed equilibrium. In addition, in our simulations we observed only subtle differences in parameter estimation compared to running longer chains of 50,000 iterations. Slow parameter convergence was observed in some simulations of complex scenarios, and in some cases the parameters did not reach convergence even after 5,000 iterations. This may be due to high levels of confounding between different categories that are closely related, or to the complexity of the mixture distribution of effect sizes for certain categories of livestock (which ultimately requires more distributions to fit the wide range of effect sizes). It is therefore suggested that the sampling chain should be extended. However, as the number of iterations increases, so do the computational and storage requirements, making it impractical to handle large cohorts at the sequence level. In summary, when computational resources are plentiful and the amount of data to be processed is limited, it is recommended to use a longer chain, as this may provide more accurate estimates. However, in cases where computational resources are limited, the use of 5,000 iterations in bovine studies still provides reasonable estimates compared to those obtained from longer chains.

6.5 Heritability partitioning for complex traits in cattle

In humans, the genetic architecture of height and disease has been extensively studied using heritability partitioning across chromosomes (Yang et al., 2011b), genomic locations (Loh et al., 2015a), and functional annotations (Finucane et al., 2015; Gusev et al., 2014; Patxot et al., 2021; Zhang et al., 2021). Studies have shown significant enrichment of heritability for quantitative traits, such as complex diseases, in regulatory elements, coding sequences, and conserved regions compared to intergenic or intronic regions (Finucane et al., 2015; Gusev et al., 2014; Patxot et al., 2021; Zhang et al., 2021). In cattle, heritability partitioning studies based on functional annotation have been less common. In addition, the functional classes used in these studies, based on gene ontology (GO) (Lingzhao et al., 2017), eQTL analyses (Xiang et al., 2023) and the Ensembl database (Bhuiyan et al., 2018), are often defined at a lower resolution than in humans. Furthermore, cattle lack a comprehensive annotation catalog that is widely used across studies, making it difficult to reuse and compare results from different studies. Nevertheless, these studies provide valuable insights into the genetic architecture of complex traits in cattle and other livestock species and have been used to select markers for genomic prediction. Xiang et al (2019a) developed the FAETH score to quantify the importance of each variant based on the enrichment levels estimated for different (functional) categories, and showed that conserved sequences and eQTLs have some of the largest SNP effects (Xiang et al., 2019a). Later, Xiang et al. (2023) estimated that eQTLs account for a significant proportion of heritability, further highlighting the importance of regulatory variants in the genetic variation of complex traits. However, the two-component approaches used in their studies may be subject to bias (e.g., risk of overestimation), as demonstrated in this thesis. They then used this information to select markers with larger genetic contributions to create a custom array for genomic selection (Xiang, 2021; Xiang et al., 2021b). In the present work, using multiple annotation groups simultaneously, we found that coding variants showed the highest per-SNP heritability for muscle-related traits and size (consistently across multiple analyses). Although both studies estimate that regulatory variants have the largest contribution to genetic variance, we found higher enrichment levels per SNP for coding variants. The estimated importance of eQTLs also differs between the two studies. In Xiang et al (2023), eQTL variants were estimated to explain up to 70% of the heritability, whereas we estimated that cis-eQTLs accounted for about 10% of the heritability explained by cis-eQTLs using GREML and BayesRR-RC. These discrepancies between our results and those of Xiang et al. (2023) may be due to differences in populations, traits, and annotations used. In humans, Luke et al. (2017) found that the %SNP heritability associated with cis-eQTLs across 30 traits was 21%, while Qi et al. (2022) reported a value of approximately 10% for both cis-eQTLs and cis-sQTLs across 12 traits. Our results are also consistent with those of Gualdrón Duarte et al. (2023), who observed in an association study in the BBC population that coding variants accounted for a substantial fraction of the genetic variance. We also showed that muscle-related regulatory elements had higher enrichment levels than other regulatory elements, and the same trend was observed for eQTLs.

This means that it is important to include this information in heritability partitioning studies. As functional annotation continues to improve, thanks to resources generated by consortia such as BovReg (Moreira et al., 2022) or FAANG (Andersson et al., 2015), or as more complete catalogs of eQTLs become available from cGTEx, a standardized and well-annotated catalog similar to that available for humans may be developed. Such a functional atlas could improve our understanding of the genetic architecture of different traits beyond muscle-related traits. Matching functional annotations such as eQTL data or ATAC-Seq from the BBC itself is promising, as the catalog we used mostly consists of data not generated in the BBC. Indeed, if functional annotation is breed specific, heritability partitioning studies will yield less informative results. Currently, most of these studies in cattle use imputed genotypes rather than sequencing data, which may not accurately capture rare variants and thus affect the accuracy of parameter estimation, especially for models with MAF stratification. The combination of improved annotation with high quality genotypes could lead to more accurate and useful results. However, we urge caution in interpreting these results for cattle and other livestock species due to the large standard errors observed for both GREML and Bayesian approaches in our study.

6.6 The use of functional annotation in genomic selection

Genomic selection has been widely adopted in cattle breeding due to its higher accuracy and its ability to significantly reduce the generation interval. It has recently been implemented in BBC, where one of the main selection criteria is muscular development. A single mutation in the myostatin gene (*MSTN*) causing the double muscling phenotype has been fixed, but estimated heritability and continued response to selection indicate that genetic variation is still present in the population, providing opportunities for further genetic improvement. In my thesis, I performed genomic selection based on imputed whole genome sequence data for muscular development traits in BBC. In addition, we incorporated functional annotation information using several models to further improve the performance of genomic selection. Compared to the accuracy based on MMD arrays used in the official evaluation, we observed systematically higher prediction accuracy – approximately 0.018 and 0.016 - for Bayesian and GBLUP models, repectively, although this difference was not always statistically significant. As discussed in the introduction, there is ongoing debate about the benefits of using sequence-level data for genomic selection, especially when considering the costs. The lack of significant improvement using sequence-level genotypes suggests that markers on MMD arrays are highly efficient at tagging ungenotyped variants, likely due to the high LD patterns in cattle. Consequently, the use of full sequence data alone does not show a clear advantage in our analysis, which is consistent with some other studies. However, in a recent GWAS in BBC, the use of sequence-level data increased the power (i.e. the significance levels). It is important to note that in our comparison, the model for full sequence data only considered polygenic effects, where all variants were equally weighted. By leveraging the fact that causative variants are unevenly distributed across different functional annotations, and that full

sequencing allows us to accurately assign functional annotations to each variant, we can incorporate this information as a prior in both GBLUP and Bayesian models. This approach can help better distinguish genetic contributions from different categories, potentially improving the accuracy of genomic selection. We observed systematically higher accuracy using a GBLUP model in which variants were partitioned based on their functional annotations. Further benefits are expected by refining the annotation process. In our second annotation approach, we divided eQTL and regulatory variants based on whether they were detected in muscle, since our phenotypes of interest were muscular development traits. However, the results were almost the same, although we observed a higher per SNP heritability for the regulatory variants detected in muscle. Nevertheless, we cannot conclude that annotation tuning is ineffective. For example, the use of relevant functional annotations has improved association signals in cattle (Jiang et al., 2019). In addition, colocalization of GWAS signals and eQTLs from relevant tissues in cGTEx demonstrated the benefits of using tissue-specific catalogs (Liu et al., 2022). In our study, the limited benefits of tuning annotations may be due to the small number of these variants detected in muscle, resulting in a limited overall impact despite larger effects per SNP. Furthermore, the effect size of variants in high LD with them may be reduced, resulting in a similar overall contribution for the same segment, thus maintaining the estimated breeding value for that segment. In cases where passenger SNPs are closely linked to causal SNPs, the benefit of knowing the causal variant becomes less significant. Another explanation is that the catalog we used is currently imperfect compared to that for humans. For example, the eQTL variant catalog used here was derived from a pilot study of the cGTEx project, which used a small group of individuals with variants called from RNA-seq data. Although the overlap of eQTL and GWAS signals has been demonstrated in a tissue-relevant manner, the limited data may affect the results. In the future, with the inclusion of leading SNPs from an improved catalog (e.g., more accurate, breed-matched, and well-matched developmental stages and tissues), we may observe improved accuracy. However, lower accuracy was observed when using functional annotation with BayesRR-RC, although BayesRR-RC is generally superior to GBLUP in annotation-free models. This is inconsistent with findings in humans, where using annotation with BayesRR-RC outperformed annotation-free predictions in most cases (Orliac et al., 2022). A possible explanation is that there is a high degree of confounding between different annotation groups, and more iterations than we specified are needed to achieve convergence. Indeed, we observed some confounding between our functional categories, such as confounding between eQTLs and other regulatory variants. However, even after doubling the number of iterations, we observed similar prediction accuracies, and the computation time increased to about 10 days for our data. In future, using improved annotation or extending the annotation to include true causal variants by using a probabilistic approach to assigning variants to groups and allowing variants to be assigned to more than one group instead of hard grouping could be beneficial. BayesRCO has shown that allowing flexibility in variant grouping improves prediction (Mollandin et al., 2022). A non-significant improvement when comparing BayesRC with and without specified annotation was also reported by Xiang et al. (2021). However, using annotation with BayesRR-RC

achieved higher prediction for stature, another complex trait we studied. This suggests that no single method may be suitable for all traits. When the LDMS approach was used to group variants in the models, we did not observe LDMS stratification in heritability estimation. However, in contrast to human studies, reduced prediction accuracy was observed when LDMS models were used for genomic prediction in cattle. Nevertheless, when we pruned the variants inspired by the LDAK model, we observed that appropriate trimming is beneficial. Although not statistically significant, this improvement was systematically observed. These results obtained with LD pruning and LDMS models seem somewhat contradictory. I suspect that heritability is overestimated in some LDMS categories and underestimated in others. This could lead to an overall improvement in heritability estimation, but not in genomic prediction. Improving genomic selection in animal breeding through functional annotation remains challenging, as methods commonly borrowed from human studies often perform worse in animals. Additionally, as the number of annotation groups increases, particularly when considering LD and MAF stratification simultaneously, we must be cautious about over-parameterization when using annotations.

To gain insight in improving the routine genomic selection, where whole genome sequence is not the first priority, we selected subsets of variants based on functional annotation or LD pruning. Variant selection offers two major advantages: significant reduction in data dimensionality, thereby reducing computational burden, and allowing the use of other models. Our results show that LD pruning reduces the number of SNPs to one-fifth without compromising accuracy; in fact, it often improves performance. Interestingly, different methods yielded different levels of performance, with the BSLMM consistently outperforming others. Overall, the best results were obtained from the subset of variants identified by functional annotation. Our study highlights that marker trimming and selection are critical not only for computational efficiency, but also for improving prediction accuracy.

The objective of my thesis was to investigate whether the accuracy of genomic prediction models could be improved using whole-genome sequence data and functional annotation. This was done in a setting where all individuals were genotyped, and is therefore only an exploratory step. If such additional information is found to be valuable, further steps should be taken before application in the field. Indeed, unlike human genetic studies where all individuals are genotyped, there are large numbers of ungenotyped individuals in livestock genetic evaluations. The ssGBLUP approach is able to exploit their information (phenotypes and pedigree) and combine it with the genomic information to achieve higher accuracy than a GBLUP (using only genotyped individuals), and is therefore the reference method in routine genomic evaluation models. Therefore, for practical field application, it will be important to use methods that can exploit whole-genome sequence data and functional information with a ssGBLUP approach. One possible strategy to do this is to give different weights to SNPs when building the GRM, a strategy that can be applied in both GBLUP and ssGBLUP settings. This method, also known as weighted ssGBLUP, uses information such as functional annotations or statistical results to assign weights to markers when calculating the GRM (Teissier et al., 2018). However, studies on the performance of weighted ssGBLUP have shown inconsistent results, with some reporting no significant

improvement over ssGBLUP (A. Liu et al., 2020b; Mehrban et al., 2021; Santana et al., 2023; Teissier et al., 2018). An alternative approach, featured ssGBLUP, has been proposed by extending GFBLUP to ssGBLUP (A. Liu et al., 2020b). Liu et al. 2020 compared featured ssGBLUP and weighted ssGBLUP with ssGBLUP in genomic prediction of milk and protein yield in dairy cattle, but found no significant differences in prediction reliability (A. Liu et al., 2020b). Marker preselection, also investigated in my PhD thesis, is an alternative to weighted ssGBLUP. It was investigated in ssGBLUP and slightly better results were obtained by including coding variants (Fragomeni et al., 2017) or variants in or near genes (Teng et al., 2022). As expected, all these approaches are derived from extensions of GBLUP, which means that they are likely to inherit properties from their GBLUP-based origins. Therefore, testing genomic data and functional annotations in GLUBP-based prediction models is informative about their extensions to ssGBLUP. As such, the model and annotation explored in the present study also provide further insight into their application in routine animal breeding.

Finally, an important consideration for routine field applications using whole-genome sequence data or functional annotation is the genetic correlation between all relevant traits, not only production traits but also traits related to fitness or environmental aspects. Recently, it has been shown that accelerating selection for primary traits (e.g. performance traits) with genomic selection could deteriorate correlated secondary traits for several reasons, including mismatched management, changing heritabilities and genetic correlations (Misztal and Lourenco, 2024). These aspects could be further amplified if functional information is incorporated into genomic selection, as it could accelerate further genetic gains and because more information may be available for certain traits. Such considerations are important in the Belgian Blue genomic evaluation as most of the recorded traits are performance traits and genomic selection was originally implemented only for these traits. Conversely, multiple-traits models that jointly consider more phenotypes may have the potential to better exploit functional information by using more data to estimate parameters and variant effects.

References

Abdollahi-Arpanahi, R., Morota, G., Peñagaricano, F., 2017. Predicting bull fertility using genomic data and biological information. J Dairy Sci. 100, 9656–9666. doi:10.3168/jds.2017-13288

Acloque, H., Harrison, P.W., Lakhal, W., Martin, F., Archibald, A.L., Beinat, M., et al., 2022. 550. Extensive functional genomics information from early developmental time points for pig and chicken., in: Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP). Presented at the World Congress on Genetics Applied to Livestock Production, Wageningen Academic Publishers, Rotterdam, the Netherlands, pp. 2281–2284. doi:10.3920/978-90-8686-940-4_550

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A.J., Mann, A.L., Kundu, K., et al., 2018. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet. 50, 424–431. doi:10.1038/s41588-018-0046-7

Alexandre, P.A., Naval-Sánchez, M., Menzies, M., Nguyen, L.T., Porto-Neto, L.R., Fortes, M.R.S., et al., 2021. Chromatin accessibility and regulatory vocabulary across indicine cattle tissues. Genome Biol. 22, 273. doi:10.1186/s13059-021-02489-7

Allison, Lizabeth A., Allison, Lizabeth Ann, 2008. Fundamental molecular biology, 2. print. ed. Blackwell Publ, Malden, MA.

Andersson, L., Archibald, A.L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W., et al., 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. 16, 57. doi:10.1186/s13059-015-0622-4

Arthur, T.D., Nguyen, J.P., D'Antonio-Chronowska, A., Jaureguy, J., Silva, N., Henson, B., et al., 2024. Multi-omic QTL mapping in early developmental tissues reveals phenotypic and temporal complexity of regulatory variants underlying GWAS loci. bioRxiv. 2024.04.10.588874. doi:10.1101/2024.04.10.588874

Ashurst, J.L., Collins, J.E., 2003. Gene annotation: prediction and testing. Annu Rev Genomics Hum Genet. 4, 69–88. doi:10.1146/annurev.genom.4.070802.110300

Bailey, T.L., Elkan, C., 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Mach Learn. 21, 51–80. doi:10.1007/BF00993379

Barbieri, M.M., Berger, J.O., 2004. Optimal predictive model selection. Ann. Statist. 32, 870–897. doi:10.1214/009053604000000238

Barral, A., Déjardin, J., 2023. The chromatin signatures of enhancers and their dynamic regulation. Nucleus. 14, 2160551. doi:10.1080/19491034.2022.2160551

Bell, A.C., West, A.G., Felsenfeld, G., 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell. 98, 387–396. doi:10.1016/s0092-8674(00)81967-4

Bhuiyan, M.S.A., Lim, D., Park, M., Lee, Soohyun, Kim, Y., Gondro, C., et al., 2018. Functional Partitioning of Genomic Variance and Genome-Wide Association Study for Carcass Traits in Korean Hanwoo Cattle Using Imputed Sequence Level SNP Data. Front Genet. 9.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 447, 799–816. doi:10.1038/nature05874

Blanchette, M., Tompa, M., 2002. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. Genome Res. 12, 739–748. doi:10.1101/gr.6902

Blobel, G.A., Higgs, D.R., Mitchell, J.A., Notani, D., Young, R.A., 2021. Testing the super-enhancer concept. Nat Rev Genet. 22, 749–755. doi:10.1038/s41576-021-00398-w

Boeva, V., 2016. Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. Front. Genet. 7. doi:10.3389/fgene.2016.00024

Boix, C.A., James, B.T., Park, Y.P., Meuleman, W., Kellis, M., 2021. Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature. 590, 300–307. doi:10.1038/s41586-020-03145-z

Boschiero, C., Gao, Y., Liu, M., Baldwin, R.L., Ma, L., Li, C.-J., et al., 2022. The Dynamics of Chromatin Accessibility Prompted by Butyrate-Induced Chromatin Modification in Bovine Cells. Ruminants. 2, 226–243. doi:10.3390/ruminants2020015

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., et al., 2008. High-resolution mapping and characterization of open chromatin across the genome. Cell. 132, 311–322. doi:10.1016/j.cell.2007.12.014

Bradford, H.L., Masuda, Y., VanRaden, P.M., Legarra, A., Misztal, I., 2019. Modeling missing pedigree in single-step genomic BLUP. J Dairy Sci. 102, 2336–2346. doi:10.3168/jds.2018-15434

Breen, E.J., MacLeod, I.M., Ho, P.N., Haile-Mariam, M., Pryce, J.E., Thomas, C.D., et al., 2022. BayesR3 enables fast MCMC blocked processing for largescale multi-trait genomic prediction and QTN mapping analysis. Commun Biol. 5, 661. doi:10.1038/s42003-022-03624-1

Brøndum, R.F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., et al., 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci. 98, 4107–4116. doi:10.3168/jds.2014-9005

Browning, B.L., Zhou, Y., Browning, S.R., 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 103, 338–348. doi:10.1016/j.ajhg.2018.07.015

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., et al., 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 523, 486–490. doi:10.1038/nature14590

Bulik-Sullivan, B., 2015. Relationship between LD Score and Haseman-Elston Regression. doi:10.1101/018283

Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H., Ripke, S., Yang, J., Patterson, N., et al., 2015. LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. Nat Genet. 47, 291–295. doi:10.1038/ng.3211

Cai, C., Wan, P., Wang, Hui, Cai, X., Wang, Jiabo, Chai, Z., et al., 2023. Transcriptional and open chromatin analysis of bovine skeletal muscle development by single-cell sequencing. Cell Prolif. 56, e13430. doi:10.1111/cpr.13430

Cai, X., Teng, J., Ren, D., Zhang, H., Li, J., Zhang, Z., 2022. Model Comparison of Heritability Enrichment Analysis in Livestock Population. Genes (Basel). 13, 1644. doi:10.3390/genes13091644

Campos, G. de los, Sorensen, D., Gianola, D., 2015. Genomic Heritability: What Is It? PLoS Genet. 11, e1005048. doi:10.1371/journal.pgen.1005048

Charlier, C., Coppieters, W., Rollin, F., Desmecht, D., Agerholm, J.S., Cambisano, N., et al., 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat Genet. 40, 449–454. doi:10.1038/ng.96

Charlier, C., Li, W., Harland, C., Littlejohn, M., Coppieters, W., Creagh, F., et al., 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. Genome Res. doi:10.1101/gr.207076.116

Chatterjee, S., Ahituv, N., 2017. Gene Regulatory Elements, Major Drivers of Human Disease. Annu Rev Genomics Hum Genet. 18, 45–63. doi:10.1146/annurev-genom-091416-035537

Chereji, R.V., Bryson, T.D., Henikoff, S., 2019. Quantitative MNase-seq accurately maps nucleosome occupancy levels. Genome Biol. 20, 198. doi:10.1186/s13059-019-1815-z

Cheville, N.F., 1999. Introduction to Veterinary Pathology. Wiley.

Christensen, O.F., Lund, M.S., 2010. Genomic prediction when some animals are not genotyped. Genetics Selection Evolution. 42, 2. doi:10.1186/1297-9686-42-2

Christmas, M.J., Kaplow, I.M., Genereux, D.P., Dong, M.X., Hughes, G.M., Li, X., et al., 2023. Evolutionary constraint and innovation across hundreds of placental mammals. Science. 380, eabn3943. doi:10.1126/science.abn3943

Clark, E.L., Archibald, A.L., Daetwyler, H.D., Groenen, M.A.M., Harrison, P.W., Houston, R.D., et al., 2020. From FAANG to fork: application of highly annotated genomes to improve farmed animal production. Genome Biology. 21, 285. doi:10.1186/s13059-020-02197-8

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R.F., et al., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 46, 858–865. doi:10.1038/ng.3034

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., et al., 2016. Next-generation genotype imputation service and methods. Nat Genet. 48, 1284–1287. doi:10.1038/ng.3656

de Koning, D.-J., 2016. Meuwissen et al. on Genomic Selection. Genetics. 203, 5–7. doi:10.1534/genetics.116.189795

de las Heras-Saldana, S., Lopez, B.I., Moghaddar, N., Park, W., Park, J., Chung, K.Y., et al., 2020. Use of gene expression and whole-genome sequence information to improve the accuracy of genomic prediction for carcass traits in Hanwoo cattle. Genet Sel Evol. 52, 54. doi:10.1186/s12711-020-00574-2

de los Campos, G., Gianola, D., Allison, D.B., 2010. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 11, 880–886. doi:10.1038/nrg2898

de Souza, N., 2012. The ENCODE project. Nat Methods. 9, 1046–1046. doi:10.1038/nmeth.2238

De Tavernier, J., Lips, D., Decuypere, E., Van Outryve, J., Pasquali, M., 2001. Ethical objections to Caesareans: implications on the future of the Belgian White Blue., in: Proceedings of Eursafe 2001 '"Food Savety, Food Quality and Food Ethics."' Presented at the Third Congress of the European Society for Agricultural and Food Ethics, A&Q, Polo per la Qualificazione del Sistema Agroalimentare; Milan, Florence, Italy.

Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J.L., Dermitzakis, E.T., 2019. Accurate, scalable and integrative haplotype estimation. Nat Commun. 10, 5436. doi:10.1038/s41467-019-13225-y

Do, D.N., Janss, L.L.G., Jensen, J., Kadarmideen, H.N., 2015. SNP annotation-based whole genomic prediction and selection: an application to feed efficiency and its component traits in pigs. J Anim Sci. 93, 2056–2063. doi:10.2527/jas.2014-8640

Doublet, A.-C., Croiseau, P., Fritz, S., Michenet, A., Hozé, C., Danchin-Burge, C., et al., 2019. The impact of genomic selection on genetic diversity and genetic gain in three French dairy cattle breeds. Genet Sel Evol. 51, 52. doi:10.1186/s12711-019-0495-1

Druet, T., Ahariz, N., Cambisano, N., Tamma, N., Michaux, C., Coppieters, W., et al., 2014a. Selection in action: dissecting the molecular underpinnings of the increasing muscle mass of Belgian Blue Cattle. BMC Genomics. 15, 796. doi:10.1186/1471-2164-15-796

Druet, T., Legarra, A., 2020. Theoretical and empirical comparisons of expected and realized relationships for the X-chromosome. Genet Sel Evol. 52, 50. doi:10.1186/s12711-020-00570-6

Druet, T., Macleod, I.M., Hayes, B.J., 2014b. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity. 112, 39–47. doi:10.1038/hdy.2013.13

Druet, T., Pérez-Pardal, L., Charlier, C., Gautier, M., 2013. Identification of large selective sweeps associated with major genes in cattle. Anim Genet. 44, 758–762. doi:10.1111/age.12073

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., et al., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature. 489, 57–74. doi:10.1038/nature11247

Edwards, S.M., Sørensen, I.F., Sarup, P., Mackay, T.F.C., Sørensen, P., 2016. Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in Drosophila melanogaster. Genetics. 203, 1871–1883. doi:10.1534/genetics.116.187161

Edwards, S.M., Thomsen, B., Madsen, P., Sørensen, P., 2015. Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle. Genet Sel Evol. 47, 60. doi:10.1186/s12711-015-0132-6

Ehsani, A., Janss, L., Pomp, D., Sørensen, P., 2016. Decomposing genomic variance using information from GWA, GWE and eQTL analysis. Anim Genet. 47, 165–173. doi:10.1111/age.12396

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., et al., 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 11, 446–450. doi:10.1038/nrg2809

Epstein, D.J., 2009. Cis-regulatory mutations in human disease. Brief Funct Genomic Proteomic. 8, 310–316. doi:10.1093/bfgp/elp021

Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., et al., 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 95, 4114–4129. doi:10.3168/jds.2011-5019

Ernst, J., Kellis, M., 2012. ChromHMM: automating chromatin state discovery and characterization. Nat Methods. 9, 215–216. doi:10.1038/nmeth.1906

Espinoza Pereira, K.N., Shan, J., Licht, J.D., Bennett, R.L., 2023. Histone mutations in cancer. Biochem Soc Trans. 51, 1749–1763. doi:10.1042/BST20210567

Eyre-Walker, A., 2010. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc Natl Acad Sci U S A. 107, 1752–1756. doi:10.1073/pnas.0906182107

Fang, L., Cai, W., Liu, S., Canela-Xandri, O., Gao, Y., Jiang, J., et al., 2020. Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. Genome Res. 30, 790–801. doi:10.1101/gr.250704.119

Fang, L., Liu, S., Liu, M., Kang, X., Lin, S., Li, B., et al., 2019. Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. BMC Biol. 17, 68. doi:10.1186/s12915-019-0687-8

Farnir, F., Coppieters, W., Arranz, J.J., Berzi, P., Cambisano, N., Grisart, B., et al., 2000. Extensive genome-wide linkage disequilibrium in cattle. Genome Res. 10, 220–227. doi:10.1101/gr.10.2.220

Fasquelle, C., Sartelet, A., Li, W., Dive, M., Tamma, N., Michaux, C., et al., 2009. Balancing Selection of a Frame-Shift Mutation in the MRC2 Gene Accounts for the Outbreak of

the Crooked Tail Syndrome in Belgian Blue Cattle. PLoS Genet. 5, e1000666. doi:10.1371/journal.pgen.1000666

Fedoriw, A.M., Stein, P., Svoboda, P., Schultz, R.M., Bartolomei, M.S., 2004. Transgenic RNAi Reveals Essential Function for CTCF in H19 Gene Imprinting. Science. 303, 238–240. doi:10.1126/science.1090934

Field, A., Adelman, K., 2020. Evaluating Enhancer Function and Transcription. Annu Rev Biochem. 89, 213–234. doi:10.1146/annurev-biochem-011420-095916

Filippova, G.N., Lindblom, A., Meincke, L.J., Klenova, E.M., Neiman, P.E., Collins, S.J., et al., 1998. A widely expressed transcription factor with multiple DNA sequence specificity, CTCF, is localized at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers. Genes Chromosomes Cancer. 22, 26–36.

Filippova, G.N., Qi, C.-F., Ulmer, J.E., Moore, J.M., Ward, M.D., Hu, Y.J., et al., 2002. Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter tts DNA-binding specificity. Cancer Res. 62, 48–52.

Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., et al., 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 47, 1228–1235. doi:10.1038/ng.3404

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., et al., 2019. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. BMC Biol. 17, 108. doi:10.1186/s12915-019-0726-5

Forrest, A.R.R., Kawaji, H., Rehli, M., Kenneth Baillie, J., de Hoon, M.J.L., Haberle, V., et al., 2014. A promoter-level mammalian expression atlas. Nature. 507, 462–470. doi:10.1038/nature13182

Fragomeni, B.O., Lourenco, D.A.L., Masuda, Y., Legarra, A., Misztal, I., 2017. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genet Sel Evol. 49, 59. doi:10.1186/s12711-017-0335-0

Frischknecht, M., Meuwissen, T.H.E., Bapst, B., Seefried, F.R., Flury, C., Garrick, D., et al., 2018. Short communication: Genomic prediction using imputed whole-genome sequence variants in Brown Swiss Cattle. J Dairy Sci. 101, 1292–1296. doi:10.3168/jds.2017-12890

Gao, Y., Liu, S., Baldwin VI, R.L., Connor, E.E., Cole, J.B., Ma, L., et al., 2022a. Functional annotation of regulatory elements in cattle genome reveals the roles of extracellular interaction and dynamic change of chromatin states in rumen development during weaning. Genomics. 114, 110296. doi:10.1016/j.ygeno.2022.110296

Gao, Yahui, Li, J., Cai, G., Wang, Y., Yang, W., Li, Y., et al., 2022b. Single-cell transcriptomic and chromatin accessibility analyses of dairy cattle peripheral blood mononuclear cells and their responses to lipopolysaccharide. BMC Genomics. 23, 338. doi:10.1186/s12864-022-08562-0

García-Ruiz, A., Cole, J.B., VanRaden, P.M., Wiggans, G.R., Ruiz-López, F.J., Van Tassell, C.P., 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci U S A. 113, E3995-4004. doi:10.1073/pnas.1519061113

Gautier, M., Faraut, T., Moazami-Goudarzi, K., Navratil, V., Foglio, M., Grohs, C., et al., 2007. Genetic and Haplotypic Structure in 14 European and African Cattle Breeds. Genetics. 177, 1059–1070. doi:10.1534/genetics.107.075804

Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., et al., 2017. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nat Genet. 49, 1421–1427. doi:10.1038/ng.3954

Georges, M., Charlier, C., Hayes, B., 2019. Harnessing genomic information for livestock improvement. Nat Rev Genet. 20, 135–156. doi:10.1038/s41576-018-0082-2

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., Fernando, R., 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics. 183, 347–363. doi:10.1534/genetics.109.103952

Gilmour, A.R., Thompson, R., Cullis, B.R., 1995. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. Biometrics. 51, 1440–1450. doi:10.2307/2533274

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., Lieb, J.D., 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 17, 877–885. doi:10.1101/gr.5533506

Goddard, M., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 136, 245–257. doi:10.1007/s10709-008-9308-0

Goddard, M.E., Hayes, B.J., 2007. Genomic selection. J Anim Breed Genet. 124, 323–330. doi:10.1111/j.1439-0388.2007.00702.x

Grant, S.F., Reid, D.M., Blake, G., Herd, R., Fogelman, I., Ralston, S.H., 1996. Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type I alpha 1 gene. Nat Genet. 14, 203–205. doi:10.1038/ng1096-203

Grobet, L., Martin, L.J., Poncelet, D., Pirottin, D., Brouwers, B., Riquet, J., et al., 1997. A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. Nat Genet. 17, 71–74. doi:10.1038/ng0997-71

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., et al., 2019. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. Nat Genet. 51, 1060–1066. doi:10.1038/s41588-019-0424-9

GTEx Consortium, 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 369, 1318–1330. doi:10.1126/science.aaz1776

GTEx Consortium, Gte.C., 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 348, 648–660. doi:10.1126/science.1262110

Gualdrón Duarte, J.L., Gori, A.-S., Hubin, X., Lourenco, D., Charlier, C., Misztal, I., et al., 2020. Performances of Adaptive MultiBLUP, Bayesian regressions, and weighted-GBLUP approaches for genomic predictions in Belgian Blue beef cattle. BMC Genomics. 21, 545. doi:10.1186/s12864-020-06921-3

Gualdrón Duarte, J.L., Yuan, C., Gori, A.-S., Moreira, G.C.M., Takeda, H., Coppieters, W., et al., 2023. Sequenced-based GWAS for linear classification traits in Belgian Blue beef cattle reveals new coding variants in genes regulating body size in mammals. Genet Sel Evol. 55, 83. doi:10.1186/s12711-023-00857-4

Guan, D., Bai, Z., Zhu, X., Zhong, C., Hou, Y., The ChickenGTEx Consortium, et al., 2023. The ChickenGTEx pilot analysis: a reference of regulatory variants across 28 chicken tissues. doi:10.1101/2023.06.27.546670

Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., et al., 2014. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 95, 535–552. doi:10.1016/j.ajhg.2014.10.004

Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J., 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 12, 186. doi:10.1186/1471-2105-12-186

Halstead, M.M., Kern, C., Saelao, P., Wang, Y., Chanthavixay, G., Medrano, J.F., et al., 2020a. A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. BMC Genomics. 21, 698. doi:10.1186/s12864-020-07078-9

Halstead, M.M., Ma, X., Zhou, C., Schultz, R.M., Ross, P.J., 2020b. Chromatin remodeling in bovine embryos indicates species-specific regulation of genome activation. Nat Commun. 11, 4654. doi:10.1038/s41467-020-18508-3

Hanset, R., 2004. Emergence and Selection of the Belgian Blue Breed., in: In Honour of the Danish B.B. Herd-Book at the Occasion of Its 25th Anniversary.

Harris, M.B., Mostecki, J., Rothman, P.B., 2005. Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function. J Biol Chem. 280, 13114–13121. doi:10.1074/jbc.M412649200

Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J., Goddard, M.E., 2010. Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. PLoS Genet. 6, e1001139. doi:10.1371/journal.pgen.1001139

Hayes, B.J., Visscher, P.M., McPartlan, H.C., Goddard, M.E., 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13, 635–643. doi:10.1101/gr.387103

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., et al., 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell. 38, 576–589. doi:10.1016/j.molcel.2010.05.004

Hertz, G.Z., Stormo, G.D., 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 15, 563–577. doi:10.1093/bioinformatics/15.7.563

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., et al., 2013. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 41, 827–841. doi:10.1093/nar/gks1284

Hoflack, G., Van den Broeck, W., Maes, D., Van Damme, K., Opsomer, G., Duchateau, L., et al., 2008. Testicular dysfunction is responsible for low sperm quality in Belgian Blue bulls. Theriogenology. 69, 323–332. doi:10.1016/j.theriogenology.2007.09.034

Holwerda, S.J.B., de Laat, W., 2013. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. Philos Trans R Soc Lond B Biol Sci. 368, 20120369. doi:10.1098/rstb.2012.0369

Hu, Y., Jiang, Z., Chen, K., Zhou, Z., Zhou, X., Wang, Y., et al., 2023. scNanoATAC-seq: a long-read single-cell ATAC sequencing method to detect chromatin accessibility and genetic variants simultaneously within an individual cell. Cell Res. 33, 83–86. doi:10.1038/s41422-022-00730-x

Huang, Y., Zhang, J., Li, X., Wu, Z., Xie, G., Wang, Yong, et al., 2023. Chromatin accessibility memory of donor cells disrupts bovine somatic cell nuclear transfer blastocysts development. The FASEB Journal. 37, e23111. doi:10.1096/fj.202300131RRR

Hujoel, M.L.A., Gazal, S., Hormozdiari, F., van de Geijn, B., Price, A.L., 2019. Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species. Am J Hum Genet. 104, 611–624. doi:10.1016/j.ajhg.2019.02.008

Hulsegge, B., Calus, M.P.L., Windig, J.J., Hoving-Bolink, A.H., Maurice-van Eijndhoven, M.H.T., Hiemstra, S.J., 2013. Selection of SNP from 50K and 777K arrays to predict breed of origin in cattle. J Anim Sci. 91, 5128–5134. doi:10.2527/jas.2013-6678

Ibáñez-Escriche, N., Forni, S., Noguera, J.L., Varona, L., 2014. Genomic information in pig breeding: Science meets industry needs. Livest Sci, Genomics Applied to Livestock Production. 166, 94–100. doi:10.1016/j.livsci.2014.05.020

Inovéo, 2020. Génomique Blanc Bleu Belge. Elevage Wallonie: Edition spéciale. 26.

Jensen, J., Su, G., Madsen, P., 2012. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC Genet. 13, 44. doi:10.1186/1471-2156-13-44

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

Johnston, D., Kim, J., Taylor, J.F., Earley, B., McCabe, M.S., Lemon, K., et al., 2021. ATAC-Seq identifies regions of open chromatin in the bronchial lymph nodes of dairy calves experimentally challenged with bovine respiratory syncytial virus. BMC Genomics. 22, 14. doi:10.1186/s12864-020-07268-5

Kadri, N.K., Zhang, J., Oget-Ebrad, C., Wang, Y., Couldrey, C., Spelman, R., et al., 2022. High male specific contribution of the X-chromosome to individual global recombination rate in dairy cattle. BMC Genomics. 23, 114. doi:10.1186/s12864-022-08328-8

Kaiser, V.B., Talmane, L., Kumar, Y., Semple, F., MacLennan, M., Deciphering Developmental Disorders Study, et al., 2021. Mutational bias in spermatogonia impacts the anatomy of regulatory sites in the human genome. Genome Res. 31, 1994–2007. doi:10.1101/gr.275407.121

Karim, L., Takeda, H., Lin, L., Druet, T., Arias, J.A.C., Baurain, D., et al., 2011. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nat Genet. 43, 405–413. doi:10.1038/ng.814

Kemper, K.E., Goddard, M.E., 2012. Understanding and predicting complex traits: knowledge from cattle. Hum Mol Genet. 21, R45-51. doi:10.1093/hmg/dds332

Kern, C., Wang, Y., Xu, X., Pan, Z., Halstead, M., Chanthavixay, G., et al., 2021. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nat Commun. 12, 1821. doi:10.1038/s41467-021-22100-8

Khansefid, M., Goddard, M.E., Haile-Mariam, M., Konstantinov, K.V., Schrooten, C., de Jong, G., et al., 2020. Improving Genomic Prediction of Crossbred and Purebred Dairy Cattle. Front. Genet. 11. doi:10.3389/fgene.2020.598580

Kim, S.S., Truong, B., Jagadeesh, K., Dey, K.K., Shen, A.Z., Raychaudhuri, S., et al., 2024. Leveraging single-cell ATAC-seq and RNA-seq to identify disease-critical fetal and adult brain cell types. Nat Commun. 15, 563. doi:10.1038/s41467-024-44742-0

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., et al., 2007. Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. Cell. 128, 1231–1245. doi:10.1016/j.cell.2006.12.048

Kimura, H., 2013. Histone modifications for human epigenome analysis. J Hum Genet. 58, 439–445. doi:10.1038/jhg.2013.66

Kolkman, I., 2010. Calving problems and calving ability in the phenotypically double muscled. University of Ghent.

Kornberg, R.D., 1974. Chromatin structure: a repeating unit of histones and DNA. Science. 184, 868–871. doi:10.1126/science.184.4139.868

Koufariotis, L., Chen, Y.-P.P., Bolormaa, S., Hayes, B.J., 2014. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. BMC Genomics. 15, 436. doi:10.1186/1471-2164-15-436

Kumasaka, N., Knights, A.J., Gaffney, D.J., 2016. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet. 48, 206–213. doi:10.1038/ng.3467

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., et al., 2015. Integrative analysis of 111 reference human epigenomes. Nature. 518, 317–330. doi:10.1038/nature14248

Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., et al., 2018. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol. 36, 70–80. doi:10.1038/nbt.4038

Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., et al., 2012. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet. 44, 247–250. doi:10.1038/ng.1108

Legarra, A., Aguilar, I., Misztal, I., 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science. 92, 4656–4663. doi:10.3168/jds.2009-2061

Leman, A.R., Noguchi, E., 2013. The replication fork: understanding the eukaryotic replication machinery and the challenges to genome duplication. Genes (Basel). 4, 1–32. doi:10.3390/genes4010001

Leporcq, C., Spill, Y., Balaramane, D., Toussaint, C., Weber, M., Bardet, A.F., 2020. TFmotifView: a webserver for the visualization of transcription factor motifs in genomic regions. Nucleic Acids Research. 48, W208–W217. doi:10.1093/nar/gkaa252

Leroy, G., 2014. Inbreeding depression in livestock species: review and meta-analysis. Anim Genet. 45, 618–628. doi:10.1111/age.12178

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., et al., 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet. 12, 1725–1735. doi:10.1093/hmg/ddg180

Leung, J.Y., McKenzie, F.E., Uglialoro, A.M., Flores-Villanueva, P.O., Sorkin, B.C., Yunis, E.J., et al., 2000. Identification of phylogenetic footprints in primate tumor necrosis factor-alpha promoters. Proc Natl Acad Sci U S A. 97, 6614–6618. doi:10.1073/pnas.97.12.6614

Li, L., Davie, J.R., 2010. The role of Sp1 and Sp3 in normal and cancer cell biology. Ann Anat. 192, 275–283. doi:10.1016/j.aanat.2010.07.010

Li, Y., Tollefsbol, T.O., 2011. DNA methylation detection: Bisulfite genomic sequencing analysis. Methods Mol Biol. 791, 11–21. doi:10.1007/978-1-61779-316-5_2

Lim, F., Ryan, G.E., Le, S.H., Solvason, J.J., Steffen, P., Farley, E.K., 2024. Affinity-optimizing variants within the ZRS enhancer disrupt limb development. Nature. 626, 151–159. doi:doi.org/10.1038/s41586-023-06922-8

Lingzhao, F., Goutam, S., Peipei, M., Guosheng, S., Ying, Y., Shengli, Z., et al., 2017. Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds. BMC Genomics. 18, 604. doi:10.1186/s12864-017-4004-z

Liu, A., Lund, M.S., Boichard, D., Karaman, E., Fritz, S., Aamand, G.P., et al., 2020a. Improvement of genomic prediction by integrating additional single nucleotide polymorphisms selected from imputed whole genome sequencing data. Heredity. 124, 37–49. doi:10.1038/s41437-019-0246-7

Liu, A., Lund, M.S., Boichard, D., Karaman, E., Guldbrandtsen, B., Fritz, S., et al., 2020b. Weighted single-step genomic best linear unbiased prediction integrating variants selected from sequencing data by association and bioinformatics analyses. Genet Sel Evol. 52, 48. doi:10.1186/s12711-020-00568-0

Liu, S., Gao, Y., Canela-Xandri, O., Wang, S., Yu, Y., Cai, W., et al., 2022. A multi-tissue atlas of regulatory variants in cattle. Nat Genet. 54, 1438–1447. doi:10.1038/s41588-022-01153-5

Liu, S., Yu, Y., Zhang, S., Cole, J.B., Tenesa, A., Wang, T., et al., 2020. Epigenomics and genotype-phenotype association analyses reveal conserved genetic architecture of

complex traits in cattle and human. BMC Biol. 18, 80. doi:10.1186/s12915-020-00792-6

Liu, X., Finucane, H.K., Gusev, A., Bhatia, G., Gazal, S., O'Connor, L., et al., 2017. Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues. Am J Hum Genet. 100, 605–616. doi:10.1016/j.ajhg.2017.03.002

Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., et al., 2015a. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet. 47, 1385–1392. doi:10.1038/ng.3431

Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., et al., 2015b. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 47, 284–290. doi:10.1038/ng.3190

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., et al., 2000. Identification of a Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-Species Sequence Comparisons. Science. 288, 136–140. doi:10.1126/science.288.5463.136

Lopez, B.I.M., An, N., Srikanth, K., Lee, S., Oh, J.-D., Shin, D.-H., et al., 2021. Genomic Prediction Based on SNP Functional Annotation Using Imputed Whole-Genome Sequence Data in Korean Hanwoo Cattle. Front Genet. 11. doi:10.3389/fgene.2020.603822

Luquette, L.J., Miller, M.B., Zhou, Z., Bohrson, C.L., Zhao, Y., Jin, H., et al., 2022. Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. Nat Genet. 54, 1564–1571. doi:10.1038/s41588-022-01180-2

MacLeod, I.M., Bowman, P.J., Vander Jagt, C.J., Haile-Mariam, M., Kemper, K.E., Chamberlain, A.J., et al., 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 17, 144. doi:10.1186/s12864-016-2443-6

MacLeod, I.M., Hayes, B.J., Goddard, M.E., 2014. The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data. Genetics. 198, 1671–1684. doi:10.1534/genetics.114.168344

MacLeod, I.M., Larkin, D.M., Lewin, H.A., Hayes, B.J., Goddard, M.E., 2013. Inferring Demography from Runs of Homozygosity in Whole-Genome Sequence, with Correction for Sequence Errors. Mol Biol Evol. 30, 2209–2223. doi:10.1093/molbev/mst125

Manen, J.-F., Savolainen, V., Simone, P., 1994. The atpB and rbcL promoters in plastid DNAs of a wide dicot range. J Mol Evol. 38, 577–582. doi:10.1007/BF00175877

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., et al., 2017. Rare and low-frequency coding variants alter human adult height. Nature. 542, 186–190. doi:10.1038/nature21039

Márquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S.S., Furlotte, N., Auton, A., et al., 2021. Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. Nat Commun. 12, 6052. doi:10.1038/s41467-021-25171-9

Maston, G.A., Evans, S.K., Green, M.R., 2006. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet. 7, 29–59. doi:10.1146/annurev.genom.7.080505.115623

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., et al., 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17, 122. doi:10.1186/s13059-016-0974-4

Mehrban, H., Naserkheil, M., Lee, D.H., Cho, C., Choi, T., Park, M., et al., 2021. Genomic Prediction Using Alternative Strategies of Weighted Single-Step Genomic BLUP for Yearling Weight and Carcass Traits in Hanwoo Beef Cattle. Genes. 12, 266. doi:10.3390/genes12020266

Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., et al., 2020. Index and biological spectrum of human DNase I hypersensitive sites. Nature. 584, 244–251. doi:10.1038/s41586-020-2559-3

Meuwissen, T., Eikje, L.S., Gjuvsland, A.B., 2024. GWABLUP: genome-wide association assisted best linear unbiased prediction of genetic values. Genet Sel Evol. 56, 17. doi:10.1186/s12711-024-00881-y

Meuwissen, T., Goddard, M., 2010. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. Genetics. 185, 623–631. doi:10.1534/genetics.110.116590

Meuwissen, T., Hayes, B., Goddard, M., 2016. Genomic selection: A paradigm shift in animal breeding. Anim Front. 6, 6–14. doi:10.2527/af.2016-0002

Meuwissen, T., Hayes, B.J., Goddard, M., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics. 157, 1819–1829. doi:10.1093/genetics/157.4.1819

Meuwissen, T., van den Berg, I., Goddard, M., 2021. On the use of whole-genome sequence data for across-breed genomic prediction and fine-scale mapping of QTL. Genet Sel Evol. 53, 19. doi:10.1186/s12711-021-00607-4

Mezger, A., Klemm, S., Mann, I., Brower, K., Mir, A., Bostick, M., et al., 2018. High-throughput chromatin accessibility profiling at single-cell resolution. Nat Commun. 9, 3647. doi:10.1038/s41467-018-05887-x

Ming, H., Sun, J., Pasquariello, R., Gatenby, L., Herrick, J.R., Yuan, Y., et al., 2021. The landscape of accessible chromatin in bovine oocytes and early embryos. Epigenetics. 16, 300–312. doi:10.1080/15592294.2020.1795602

Misztal, I., 2008. Reliable computing in estimation of variance components. Journal of Animal Breeding and Genetics. 125, 363–370. doi:10.1111/j.1439-0388.2008.00774.x

Misztal, I., Lourenco, D., 2024. Potential negative effects of genomic selection. Journal of Animal Science. 102, skae155. doi:10.1093/jas/skae155

Molineris, I., Grassi, E., Ala, U., Di Cunto, F., Provero, P., 2011. Evolution of promoter affinity for transcription factors in the human lineage. Mol Biol Evol. 28, 2173–2183. doi:10.1093/molbev/msr027

Mollandin, F., Gilbert, H., Croiseau, P., Rau, A., 2022. Accounting for overlapping annotations in genomic prediction models of complex traits. BMC Bioinformatics. 23, 365. doi:10.1186/s12859-022-04914-5

Monroe, J.G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., et al., 2022. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature. 602, 101–105. doi:10.1038/s41586-021-04269-6

Moon, H., Filippova, G., Loukinov, D., Pugacheva, E., Chen, Q., Smith, S.T., et al., 2005. CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. EMBO Rep. 6, 165–170. doi:10.1038/sj.embor.7400334

Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., et al., 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 583, 699–710. doi:10.1038/s41586-020-2493-4

Moore, L.D., Le, T., Fan, G., 2013. DNA Methylation and Its Basic Function. Neuropsychopharmacol. 38, 23–38. doi:10.1038/npp.2012.112

Moreira, G.C.M., Dupont, S., Becker, D., Salavati, M., Clark, R., Clark, E. l., et al., 2022. 545. Multi-dimensional functional annotation of bovine genome for the BovReg project., in:

Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP). Wageningen Academic Publishers, pp. 2261–2264. doi:10.3920/978-90-8686-940-4_545

Morova, T., Ding, Y., Huang, C.-C.F., Sar, F., Schwarz, T., Giambartolomei, C., et al., 2022. Optimized high-throughput screening of non-coding variants identified from genome-wide association studies. Nucleic Acids Res. 51, e18. doi:10.1093/nar/gkac1198

Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R., Visscher, P.M., 2015. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. PLOS Genet. 11, e1004969. doi:10.1371/journal.pgen.1004969

Nagy, G., Nagy, L., 2020. Motif grammar: The basis of the language of gene expression. Comput Struct Biotechnol J. 18, 2026. doi:10.1016/j.csbj.2020.07.007

Nagy, P.L., Cleary, M.L., Brown, P.O., Lieb, J.D., 2003. Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. Proc Natl Acad Sci U S A. 100, 6364–6369. doi:10.1073/pnas.1131966100

Nayee, N., Sahana, G., Gajjar, S., Sudhakar, A., Trivedi, K., Lund, M.S., et al., 2018. Suitability of existing commercial single nucleotide polymorphism chips for genomic studies in Bos indicus cattle breeds and their Bos taurus crosses. J Anim Breed Genet. 135, 432–441. doi:10.1111/jbg.12356

Neyret-Kahn, H., Fontugne, J., Meng, X.Y., Groeneveld, C.S., Cabel, L., Ye, T., et al., 2023. Epigenomic mapping identifies an enhancer repertoire that regulates cell identity in bladder cancer through distinct transcription factor networks. Oncogene. 42, 1524–1542. doi:10.1038/s41388-023-02662-1

Nicolazzi, E.L., Picciolini, M., Strozzi, F., Schnabel, R.D., Lawley, C., Pirani, A., et al., 2014. SNPchiMp: a database to disentangle the SNPchip jungle in bovine livestock. BMC Genomics. 15, 123. doi:10.1186/1471-2164-15-123

Niu, Q., Zhang, T., Xu, Ling, Wang, T., Wang, Z., Zhu, B., et al., 2021. Integration of selection signatures and multi-trait GWAS reveals polygenic genetic architecture of carcass traits in beef cattle. Genomics. 113, 3325–3336. doi:10.1016/j.ygeno.2021.07.025

O'Connor, L.J., Gusev, A., Liu, X., Loh, P.-R., Finucane, H.K., Price, A.L., 2017. Estimating the proportion of disease heritability mediated by gene expression levels. doi:10.1101/118018

Oomen, M.E., Hansen, A.S., Liu, Y., Darzacq, X., Dekker, J., 2019. CTCF sites display cell cycle–dependent dynamics in factor binding and nucleosome positioning. Genome Res. 29, 236–249. doi:10.1101/gr.241547.118

Orliac, E.J., Trejo Banos, D., Ojavee, S.E., Läll, K., Mägi, R., Visscher, P.M., et al., 2022. Improving GWAS discovery and genomic prediction accuracy in biobank data. Proc Natl Acad Sci U S A. 119, e2121279119. doi:10.1073/pnas.2121279119

Panigrahi, A., O'Malley, B.W., 2021. Mechanisms of enhancer action: the known and the unknown. Genome Biol. 22, 108. doi:10.1186/s13059-021-02322-1

Pareek, C.S., Smoczynski, R., Pierzchala, M., Czarnik, U., Tretyn, A., 2011. From genotype to phenotype in bovine functional genomics. Brief Funct Genomics. 10, 165–171. doi:10.1093/bfgp/elr019

Park, P.J., 2009. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet. 10, 669–680. doi:10.1038/nrg2641

Parmar, J.J., Padinhateeri, R., 2020. Nucleosome positioning and chromatin organization. Current Opinion in Structural Biology. 64, 111–118. doi:10.1016/j.sbi.2020.06.021

Patxot, M., Banos, D.T., Kousathanas, A., Orliac, E.J., Ojavee, S.E., Moser, G., et al., 2021. Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits. Nat Commun. 12, 6972. doi:10.1038/s41467-021-27258-9

Pazokitoroudi, A., Wu, Y., Burch, K.S., Hou, K., Zhou, A., Pasaniuc, B., et al., 2020. Efficient variance components analysis across millions of genomes. Nat Commun. 11, 4020. doi:10.1038/s41467-020-17576-9

Pennisi, E., 2012. ENCODE Project Writes Eulogy for Junk DNA. Science. 337, 1159–1161. doi:10.1126/science.337.6099.1159

Pérez-Enciso, M., Rincón, J.C., Legarra, A., 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet Sel Evol. 47, 43. doi:10.1186/s12711-015-0117-5

Phillips, J.E., Corces, V.G., 2009. CTCF: master weaver of the genome. Cell. 137, 1194–1211. doi:10.1016/j.cell.2009.06.001

Pimentel, E. da C.G., Erbe, M., Koenig, S., Simianer, H., 2011. Genome Partitioning of Genetic Variation for Milk Production and Composition Traits in Holstein Cattle. Front. Genet. 2. doi:10.3389/fgene.2011.00019

Pocrnic, I., Lourenco, D., Misztal, I., 2024. Single nucleotide polymorphism profile for quantitative trait nucleotide in populations with small effective size and its impact on mapping and genomic predictions. Genetics. 227, iyae103. doi:10.1093/genetics/iyae103

Powell, J., Talenti, A., Fisch, A., Hemmink, J.D., Paxton, E., Toye, P., et al., 2023. Profiling the immune epigenome across global cattle breeds. Genome Biology. 24, 127. doi:10.1186/s13059-023-02964-3

Qi, T., Wu, Y., Fang, H., Zhang, F., Liu, S., Zeng, J., et al., 2022. Genetic control of RNA splicing and its distinct role in complex trait variation. Nat Genet. 54, 1355–1363. doi:10.1038/s41588-022-01154-4

Quackenbush, J., 2022. Looking back at the first twenty years of genomics. Quant Biol. 10, 6–16. doi:10.15302/J-QB-021-0286

Quick, C., Anugu, P., Musani, S., Weiss, S.T., Burchard, E.G., White, M.J., et al., 2020. Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse populations. Genet Epidemiol. 44, 537–549. doi:10.1002/gepi.22326

Raymond, B., Bouwman, A.C., Schrooten, C., Houwing-Duistermaat, J., Veerkamp, R.F., 2018. Utility of whole-genome sequence data for across-breed genomic prediction. Genet Sel Evol. 50, 27. doi:10.1186/s12711-018-0396-8

Reijns, M.A.M., Kemp, H., Ding, J., de Procé, S.M., Jackson, A.P., Taylor, M.S., 2015. Lagging-strand replication shapes the mutational landscape of the genome. Nature. 518, 502–506. doi:10.1038/nature14183

Richmond, T.J., Davey, C.A., 2003. The structure of DNA in the nucleosome core. Nature. 423, 145–150. doi:10.1038/nature01595

Rincon, G., Weber, K.L., Eenennaam, A.L.V., Golden, B.L., Medrano, J.F., 2011. Hot topic: performance of bovine high-density genotyping platforms in Holsteins and Jerseys. J Dairy Sci. 94, 6116–6121. doi:10.3168/jds.2011-4764

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., et al., 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 4, 651–657. doi:10.1038/nmeth1068

Robinson, M.R., Santure, A.W., DeCauwer, I., Sheldon, B.C., Slate, J., 2013. Partitioning of genetic variation across the genome using multimarker methods in a wild bird population. Mol Ecol. 22, 3963–3980. doi:10.1111/mec.12375

Ros-Freixedes, R., Johnsson, M., Whalen, A., Chen, C.-Y., Valente, B.D., Herring, W.O., et al., 2022. Genomic prediction with whole-genome sequence data in intensely selected pig lines. Genet Sel Evol. 54, 65. doi:10.1186/s12711-022-00756-0

Rotem, A., Ram, O., Shoresh, N., Sperling, R.A., Goren, A., Weitz, D.A., et al., 2015. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat Biotechnol. 33, 1165–1172. doi:10.1038/nbt.3383

Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., et al., 2010. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science. 330, 1787–1797. doi:10.1126/science.1198374

Rubinacci, S., Delaneau, O., Marchini, J., 2020. Genotype imputation using the Positional Burrows Wheeler Transform. PLoS Genet. 16, e1009049. doi:10.1371/journal.pgen.1009049

Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., López-Bigas, N., 2016. Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature. 532, 264–267. doi:10.1038/nature17661

Salavati, M., Clark, R., Becker, D., Kühn, C., Plastow, G., Dupont, S., et al., 2023. Improving the annotation of the cattle genome by annotating transcription start sites in a diverse set of tissues and populations using Cap Analysis Gene Expression sequencing. G3 (Bethesda). 13, jkad108. doi:10.1093/g3journal/jkad108

Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W., Lenhard, B., 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 32, D91–D94. doi:10.1093/nar/gkh012

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

Sartelet, A., Druet, T., Michaux, C., Fasquelle, C., Géron, S., Tamma, N., et al., 2012. A Splice Site Variant in the Bovine RNF11 Gene Compromises Growth and Regulation of the Inflammatory Response. PLOS Genet. 8, e1002581. doi:10.1371/journal.pgen.1002581

Sartelet, A., Li, W., Pailhoux, E., Richard, C., Tamma, N., Karim, L., et al., 2015. Genome-wide next-generation DNA and RNA sequencing reveals a mutation that perturbs splicing of the phosphatidylinositol glycan anchor biosynthesis class H gene (PIGH) and causes arthrogryposis in Belgian Blue cattle. BMC Genomics. 16, 316. doi:10.1186/s12864-015-1528-y

Sartelet, A., Stauber, T., Coppieters, W., Ludwig, C.F., Fasquelle, C., Druet, T., et al., 2014. A missense mutation accelerating the gating of the lysosomal Cl−/H+-exchanger ClC-7/Ostm1 causes osteopetrosis with gingival hamartomas in cattle. Dis Model Mech. 7, 119–128. doi:10.1242/dmm.012500

Sarup, P., Jensen, J., Ostersen, T., Henryon, M., Sørensen, P., 2016. Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. BMC Genetics. 17, 11. doi:10.1186/s12863-015-0322-9

Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., et al., 2019. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat Biotechnol. 37, 925–936. doi:10.1038/s41587-019-0206-z

Sethi, A., Gu, M., Gumusgoz, E., Chan, L., Yan, K.-K., Rozowsky, J., et al., 2020. Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. Nat Methods. 17, 807–814. doi:10.1038/s41592-020-0907-8

Snedeker, J., Wooten, M., Chen, X., 2017. The Inherent Asymmetry of DNA Replication. Annu Rev Cell Dev Biol. 33, 291–318. doi:10.1146/annurev-cellbio-100616-060447

Solé, M., Gori, A.-S., Faux, P., Bertrand, A., Farnir, F., Gautier, M., et al., 2017. Age-based partitioning of individual genomic inbreeding levels in Belgian Blue cattle. Genet Sel Evol. 49, 92. doi:10.1186/s12711-017-0370-x

Sørensen, P., Edwards, S.M., Rohde, P.D., 2014. Genomic feature model., in: 10th World Congress on Genetics Applied to Livestock Production. Presented at the Congress on Genetics Applied to Livestock Production (WCGALP), Vancouver, Canada.

Speed, D., Balding, D.J., 2019. SumHer better estimates the SNP heritability of complex traits from summary statistics. Nat Genet. 51, 277–284. doi:10.1038/s41588-018-0279-5

Speed, D., Balding, D.J., 2015. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 16, 33–44. doi:10.1038/nrg3821

Speed, D., Balding, D.J., 2014. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 24, 1550–1557. doi:10.1101/gr.169375.113

Speed, D., Cai, N., Johnson, M.R., Nejentsev, S., Balding, D.J., 2017. Reevaluation of SNP heritability in complex human traits. Nat Genet. 49, 986–992. doi:10.1038/ng.3865

Speed, D., Hemani, G., Johnson, M.R., Balding, D.J., 2012. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 91, 1011–1021. doi:10.1016/j.ajhg.2012.10.010

Speed, D., Holmes, J., Balding, D.J., 2020. Evaluating and improving heritability models using summary statistics. Nat Genet. 52, 458–462. doi:10.1038/s41588-020-0600-y

spicuglia, salvatore, Vanhille, L., 2012. Chromatin signatures of active enhancers. Nucleus. 3, 126–131. doi:10.4161/nucl.19232

Srinivasan, L., Atchison, M.L., 2004. YY1 DNA binding and PcG recruitment requires CtBP. Genes Dev. 18, 2596–2601. doi:10.1101/gad.1228204

Staden, R., 1984. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. 12, 505–519. doi:10.1093/nar/12.1part2.505

Strandén, I., Christensen, O.F., 2011. Allele coding in genomic evaluation. Genetics Selection Evolution. 43, 25. doi:10.1186/1297-9686-43-25

Strandén, I., Garrick, D.J., 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci. 92, 2971–2975. doi:10.3168/jds.2008-1929

Stunnenberg, H.G., Abrignani, S., Adams, D., Almeida, M. de, Altucci, L., Amin, V., et al., 2016. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell. 167, 1145–1149. doi:10.1016/j.cell.2016.11.007

Su, G., Christensen, O.F., Janss, L., Lund, M.S., 2014. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J Dairy Sci. 97, 6547–6559. doi:10.3168/jds.2014-8210

Sullivan, P.F., Meadows, J.R.S., Gazal, S., Phan, B.N., Li, X., Genereux, D.P., et al., 2023. Leveraging base-pair mammalian constraint to understand genetic variation and human disease. Science. 380, eabn2937. doi:10.1126/science.abn2937

Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., Jones, R.T., 1988. Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*). J Mol Biol. 203, 439–455. doi:10.1016/0022-2836(88)90011-3

Teissier, M., Larroque, H., Robert-Granié, C., 2018. Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: a quantitative trait influenced by a major gene. Genet Sel Evol. 50, 31. doi:10.1186/s12711-018-0400-3

Teng, J., Gao, Y., Yin, H., Bai, Z., Liu, S., Zeng, H., et al., 2024. A compendium of genetic regulatory effects across pig tissues. Nat Genet. 56, 112–123. doi:10.1038/s41588-023-01585-7

Teng, Ye, S., Ning, G.A.O., Chen, Z., Diao, S., Li, X., et al., 2022. Incorporating genomic annotation into single-step genomic prediction with imputed whole-genome sequence data. Journal of Integrative Agriculture. 21, 1126–1136.

Tost, J., 2009. DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. Methods Mol Biol. 507, 3–20. doi:10.1007/978-1-59745-522-0_1

Trynka, G., Westra, H.-J., Slowikowski, K., Hu, X., Xu, H., Stranger, B.E., et al., 2015. Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. Am J Hum Genet. 97, 139–152. doi:10.1016/j.ajhg.2015.05.016

Tsompana, M., Buck, M.J., 2014. Chromatin accessibility: a window into the genome. Epigenetics & Chromatin. 7, 33. doi:10.1186/1756-8935-7-33

van Binsbergen, R., Calus, M.P.L., Bink, M.C.A.M., van Eeuwijk, F.A., Schrooten, C., Veerkamp, R.F., 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 47, 71. doi:10.1186/s12711-015-0149-x

van de Geijn, B., Finucane, H., Gazal, S., Hormozdiari, F., Amariuta, T., Liu, X., et al., 2020. Annotations capturing cell type-specific TF binding explain a large fraction of disease heritability. Hum Mol Genet. 29, 1057–1067. doi:10.1093/hmg/ddz226

Van Eenennaam, A.L., Weigel, K.A., Young, A.E., Cleveland, M.A., Dekkers, J.C.M., 2014. Applied animal genomics: results from the field. Annu Rev Anim Biosci. 2, 105–139. doi:10.1146/annurev-animal-022513-114119

Van Laere, A.-S., Nguyen, M., Braunschweig, M., Nezer, C., Collette, C., Moreau, L., et al., 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature. 425, 832–836. doi:10.1038/nature02064

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J Dairy Sci. 91, 4414–4423. doi:10.3168/jds.2007-0980

VanRaden, P.M., Tooker, M.E., O'Connell, J.R., Cole, J.B., Bickhart, D.M., 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol. 49, 32. doi:10.1186/s12711-017-0307-4

Veerkamp, R.F., Bouwman, A.C., Schrooten, C., Calus, M.P.L., 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. Genet Sel Evol. 48, 95. doi:10.1186/s12711-016-0274-1

Visscher, P.M., Hill, W.G., Wray, N.R., 2008. Heritability in the genomics era — concepts and misconceptions. Nat Rev Genet. 9, 255–266. doi:10.1038/nrg2322

Vos, J. de, Derks, M.F.L., Kurylo, C., Groenen, M.A.M., Madsen, O., 2023. GSM-pipeline: GENE-SWitCH pipeline for comprehensive bisulfite sequencing analysis. doi:10.21203/rs.3.rs-2984574/v1

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., et al., 2021. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 53, 1300–1310. doi:10.1038/s41588-021-00913-z

Wang, X., Gao, Yahui, Li, C., Fang, L., Liu, G.E., Zhao, X., et al., 2023. The single-cell transcriptome and chromatin accessibility datasets of peripheral blood mononuclear cells in Chinese holstein cattle. BMC Genomic Data. 24, 39. doi:10.1186/s12863-023-01139-0

Wasserman, W.W., Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 5, 276–287. doi:10.1038/nrg1315

West, A.G., Gaszner, M., Felsenfeld, G., 2002. Insulators: many functions, many mechanisms. Genes Dev. 16, 271–288. doi:10.1101/gad.954702

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., et al., 2000. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. 28, 316–319.

Wray, N.R., Kemper, K.E., Hayes, B.J., Goddard, M.E., Visscher, P.M., 2019. Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. Genetics. 211, 1131–1141. doi:10.1534/genetics.119.301859

Xiang, R., 2021. Bayesian genome-wide analysis of cattle traits using variants with functional and evolutionary significance. Biorxiv. doi:https://doi.org/10.1101/2021.05.05.442705

Xiang, R., Berg, I. van den, MacLeod, I.M., Hayes, B.J., Prowse-Wilkins, C.P., Wang, M., et al., 2019a. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. Proc Natl Acad Sci U S A. 116, 19398–19408. doi:10.1073/pnas.1904159116

Xiang, R., Berg, I. van den, MacLeod, I.M., Hayes, B.J., Prowse-Wilkins, C.P., Wang, M., et al., 2019b. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. Proc Natl Acad Sci U S A. 116, 19398–19408. doi:10.1073/pnas.1904159116

Xiang, R., Breen, E.J., Prowse-Wilkins, C.P., Chamberlain, A.J., Goddard, M.E., 2021a. Bayesian genome-wide analysis of cattle traits using variants with functional and evolutionary significance. bioRxiv. 2021.05.05.442705. doi:10.1101/2021.05.05.442705

Xiang, R., Fang, L., Liu, S., Macleod, I.M., Liu, Z., Breen, E.J., et al., 2023. Gene expression and RNA splicing explain large proportions of the heritability for complex traits in cattle. Cell Genom. 3, 100385. doi:10.1016/j.xgen.2023.100385

Xiang, R., MacLeod, I.M., Daetwyler, H.D., de Jong, G., Connor, E.E., Schrooten, C., et al., 2021b. Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. Nat Commun. 12, 860. doi:10.1038/s41467-021-21001-0

Xiao, X., Xu, Z.-C., Qiu, W.-R., Wang, P., Ge, H.-T., Chou, K.-C., 2019. iPSW(2L)-PseKNC: A two-layer predictor for identifying promoters and their strength by hybrid features via pseudo K-tuple nucleotide composition. Genomics. 111, 1785–1793. doi:10.1016/j.ygeno.2018.12.001

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., et al., 2015. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 47, 1114–1120. doi:10.1038/ng.3390

Yang, J., Corces, V.G., 2011. Chromatin Insulators: A Role in Nuclear Organization and Gene Expression. Adv Cancer Res. 110, 43–76. doi:10.1016/B978-0-12-386469-7.00003-7

Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2011a. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 88, 76. doi:10.1016/j.ajhg.2010.11.011

Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., et al., 2011b. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 43, 519–525. doi:10.1038/ng.823

Yang, J., Zeng, J., Goddard, M.E., Wray, N.R., Visscher, P.M., 2017. Concepts, estimation and interpretation of SNP-based heritability. Nat Genet. 49, 1304–1310. doi:10.1038/ng.3941

Ye, S., Li, J., Zhang, Z., 2020. Multi-omics-data-assisted genomic feature markers preselection improves the accuracy of genomic prediction. J Anim Sci Biotechnol. 11, 109. doi:10.1186/s40104-020-00515-5

Yuan, C., Gualdrón Duarte, J.L., Takeda, H., Georges, M., Druet, T., 2024. Evaluation of heritability partitioning approaches in livestock populations. BMC Genomics. 25, 690. doi:10.1186/s12864-024-10600-y

Yuan, C., Tang, L., Lopdell, T., Petrov, V.A., Oget-Ebrad, C., Moreira, G.C.M., et al., 2023. An organism-wide ATAC-seq peak catalog for the bovine and its use to identify regulatory variants. Genome Res. 33, 1848–1864. doi:10.1101/gr.277947.123

Zeng, J., de Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., et al., 2018. Signatures of negative selection in the genetic architecture of human complex traits. Nat Genet. 50, 746–753. doi:10.1038/s41588-018-0101-4

Zhang, Q., Privé, F., Vilhjálmsson, B., Speed, D., 2021. Improved genetic prediction of complex traits from individual-level data or summary statistics. Nat Commun. 12, 4192. doi:10.1038/s41467-021-24485-y

Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H.P., Longin, F.H., et al., 2012. Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet. 124, 769–776. doi:10.1007/s00122-011-1745-y

Zhen, Y., Andolfatto, P., 2012. Methods to Detect Selection on Noncoding DNA. Methods Mol Biol. 856, 141–159. doi:10.1007/978-1-61779-585-5_6

Zheng, Z., Liu, S., Sidorenko, J., Wang, Y., Lin, T., Yengo, L., et al., 2024. Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. Nat Genet. 56, 767–777. doi:10.1038/s41588-024-01704-y

Zhou, X., Carbonetto, P., Stephens, M., 2013. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet. 9, e1003264. doi:10.1371/journal.pgen.1003264

Zhu, P., Schon, M., Questa, J., Nodine, M., Dean, C., 2023. Causal role of a promoter polymorphism in natural variation of the *Arabidopsis* floral repressor gene *FLC*. Curr Biol. 33, 4381-4391.e3. doi:10.1016/j.cub.2023.08.079