The conservation of human functional variants and their effects across mammals

R. Zhao[1], A.Talenti[1], L. Fang[1], S. Liu[2], G. Liu[3], N.P. Chue Hong[4], A. Tenesa[1], M. A. Hassan[1], J.G.D.Prendergast[1]*

[1] The Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, United Kingdom

[2] Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang 310024, China

[3] Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service, Agricultural Research Service, USDA, Beltsville, Maryland 20705, USA

[4] EPCC, Bayes Centre, 47 Potterrow, Edinburgh, EH8 9BT, United Kingdom

*To whom correspondence should be addressed:

(james.prendergast@roslin.ed.ac.uk).

Abstract

Despite the clear potential of livestock models of human functional variants to provide important insights into the biological mechanisms driving human diseases and traits, their use to date has been limited. Generating such models via genome editing is costly and time consuming, and it is unclear which variants will have conserved effects across species. In this study we address these issues by studying naturally occurring livestock models of human functional variants. We show that orthologues of over 1.6 million human variants are already segregating in domesticated mammalian species, including several hundred previously directly linked to human traits and diseases. Models of variants linked to particular phenotypes, including metabolomic disorders

and height, have been preferentially maintained across species, meaning studying the genetic basis of these phenotypes is particularly tractable in livestock. Using machine learning we demonstrate it is possible to identify human variants that are more likely to have an existing livestock orthologue, and, importantly, we show that the effects of functional variants are often conserved in livestock, acting on orthologous genes with the same direction of effect. Consequently, this work demonstrates the substantial potential of naturally occurring livestock carriers of orthologues of human functional variants to disentangle their functional impacts.

Introduction

Animal models are widely used across the biological sciences. From the development of vaccines and use as models of human diseases, to addressing fundamental questions about human biology. Importantly animal models provide the ability to test the effect of manipulating key variables in a controlled fashion, in ways that are not possible in human populations. For example, by altering the genome via genome editing. The introduction of variants thought to be functional in humans into animal models enables a range of studies, from the characterization of their downstream impacts on the expression of genes, to how different alleles respond to different interventions such as drug treatments.

By far the most widely used mammalian animal models are rodents, due to their ease of handling and short generation times. But rodent models have several limitations. Most importantly humans and rodents are physiologically very different, with the pathogenesis of diseases often differing substantially between the species. This has been proposed as a key driver of why less than 8% of cancer studies that are based

on animal models result in a clinical trial (Käser, 2021). Furthermore, the sizes of rodent organs poorly match those of humans, and it is difficult to serially sample rodent models due to their smaller size. Although the use of primate models can overcome many of these limitations their use is limited by both cost and ethical considerations (Käser, 2021). For these reasons livestock species have been proposed as more effective animal models in many scenarios (Meurens et al., 2012; Ziegler et al., 2016). Pigs in particular have a similar size, physiology and anatomy to humans (Walters & Prather, 2013), and have been shown to have more similar gene expression patterns to humans than rodents (Sjöstedt et al., 2020). As a result they are increasingly used in translational research, from toxicology testing of pharmaceuticals to the development of transgenic models of human diseases ranging from cystic fibrosis and diabetes to neurodegenerative disorders (Lunney et al., 2021). However, livestock models of human functional genetic variants have major drawbacks: they are expensive and time-consuming to generate. As well as the substantial time and costs associated with generating and implanting the genome edited embryos, it is necessary to maintain the mothers through long pregnancies in areas suitable for genetically modified animals, with no prospects of recouping the costs through selling the animals afterwards. There are also further ethical considerations to such transgenic projects, with the public often skeptical of the merits of artificially introducing human variants into other species.

Therefore, despite the clear merits of being able to assay the effects of human functional variants in livestock models, transgenic experiments come with several obstacles. Even among mice, the number of truly “humanized” models, i.e., where the directly orthologous mouse base or sequence has been altered to match that in

humans, is low. Traditionally transgenic mouse models involve the random insertion of transgenes into the genome, meaning they lose their wider genomic context and potential impacts on downstream functions and mechanisms (F. Zhu et al., 2019). To properly model human functional variants, the same changes need to be made at orthologous locations, with both alleles present among the animal model.

A relatively under-explored alternative to the *de novo* generation of animal models is the study of natural orthologues of human functional variants. The 1000 bulls project alone identified over 84 million cattle single nucleotide polymorphisms (Hayes & Daetwyler, 2019), meaning approximately 1 in every 32 bases in the cattle genome is polymorphic. This though is potentially an underestimate of the expected probability of a human variant having a cattle orthologue, as polymorphisms are known to be dependent on the underlying sequence. For example, CpG sites are known to be susceptible to deamination, likely raising the probability of such sites being polymorphic across species. This suggests there are potentially many natural orthologues of human functional variants, meaning the effect of these variants can be studied in large mammalian models, potentially at scale, without resorting to transgenic approaches. Supporting this idea, although rare in the literature, some examples of functional variants being found naturally across different mammalian species have already been reported. For example, a missense change linked to coat colour found segregating among both dogs and water buffalo (Dutta et al., 2020). In recent work, non-naturally occurring coding changes in mice and zebrafish were compared to these found in humans, with orthologues of human pathogenic Clinvar variants shown to more likely also to lead to a detectable phenotypic change in zebrafish than other variants (Pir et al., 2022). To date there has though been little

genome-wide study of the natural orthologues of human functional variants and the conservation of their effects across mammals. In part this has resulted from the fact that the precise functional variant underlying most human quantitative trait loci and genome-wide association loci have been unknown. However, high resolution functional datasets and fine-mapping approaches have begun to disentangle causative variants from those simply in linkage disequilibrium (Broekema et al., 2020; Schaid et al., 2018).

Studying the impact of these functional variants has the potential to inform our understanding of phenotypes beyond just humans. This is because livestock species are not only good models for humans but the reverse is also true. Substantially more biological data and insights have been generated for humans than livestock, and characterizing how human functional variants effect corresponding phenotypes in livestock may provide insights into how to improve the production and health of domesticated animals. For example, the genetic basis of stature in cattle has already been shown to have parallels of that in humans (Bouwman et al., 2018), and better understanding functional variants linked to height could provide potential avenues for adjusting livestock body size.

The aim of this study was, therefore, to characterize the extent to which natural orthologues of human variants are found in domesticated species. Using machine learning we characterize the features associated with the presence of orthologues across species, investigate the presence of functional variants linked to diseases and traits across mammals, and determine where their effects on downstream phenotypes are conserved. We highlight how orthologues of human functional variants are likely a

119 valuable resource to better understand the genetic basis of both human and livestock
120 phenotypes.

121 **Results**

122

123 ***Extensive sharing of variants across species***

124

125 To investigate how often the same variants are found across species, we compared
126 the 78 million human SNPs identified in the 1000 genomes cohort of ~2500 diverse
127 Individuals (The 1000 Genomes Project Consortium et al., 2015) to the variants
128 identified in cohorts of 477 cattle (Dutta et al., 2020) and 409 pigs (C. Li et al., 2021).
129 In total 35 and 34 million of the human variants could be mapped to an orthologous
130 location in the pig and cattle genome, respectively (Figure 1A). Of these 3.7 and 3.0%
131 overlapped an orthologous variant segregating in one of these other species, with 55.4

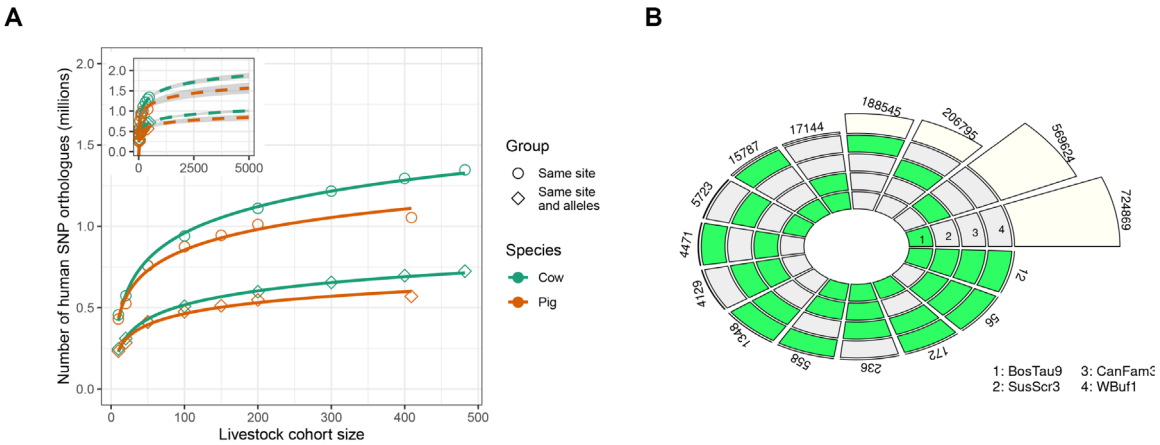


Figure 1. Frequency of variant sharing across species. (A) Number of human (1000 genomes) SNPs that have a SNP at the orthologous location in each other species. Counts are broken down into where the SNPs have the same alleles across species (same site and alleles) or simply coincide, i.e. irrespective of allele change. The inset shows the number of orthologous SNPs expected in larger cohorts when extrapolating the curves. (B) The number of human variants overlapping a variant found in one or more other species with a matching allele change.

and 55.8% of these showing the exact same allele change. Consequently over 1.1 million human variants have a direct orthologue in at least one of these two livestock cohorts. Intersecting the same human polymorphisms with variants in cohorts of two further domesticated species, 722 dogs (Plassais et al., 2019) and 81 water buffalo (Dutta et al., 2020), revealed that 1,654,806 are found in at least one of these four mammalian cohorts (Figure 1B).

The number of variants shared across cohorts from different species is expected to be a function of the number of samples in each cohort. To characterize this relationship we randomly down-sampled the pig and cattle cohorts and recalculated the observed overlap with the total set of human variants. As shown in Figure 1A the number of variants overlapping the human dataset had not plateaued for either species, suggesting larger cohorts would continue to identify even more orthologues of human variants. For example, extrapolating the results to 5000 samples in corresponding cohorts suggests over 840,000 pig and 1,000,000 cattle orthologues of human variants would potentially be detected (Figure 1A). As expected sample diversity/relatedness is also an important factor with more diverse cohorts leading to more orthologous variants being identified (Supplementary Figure 1). This suggests exact orthologues of several million human variants are naturally segregating among livestock species.

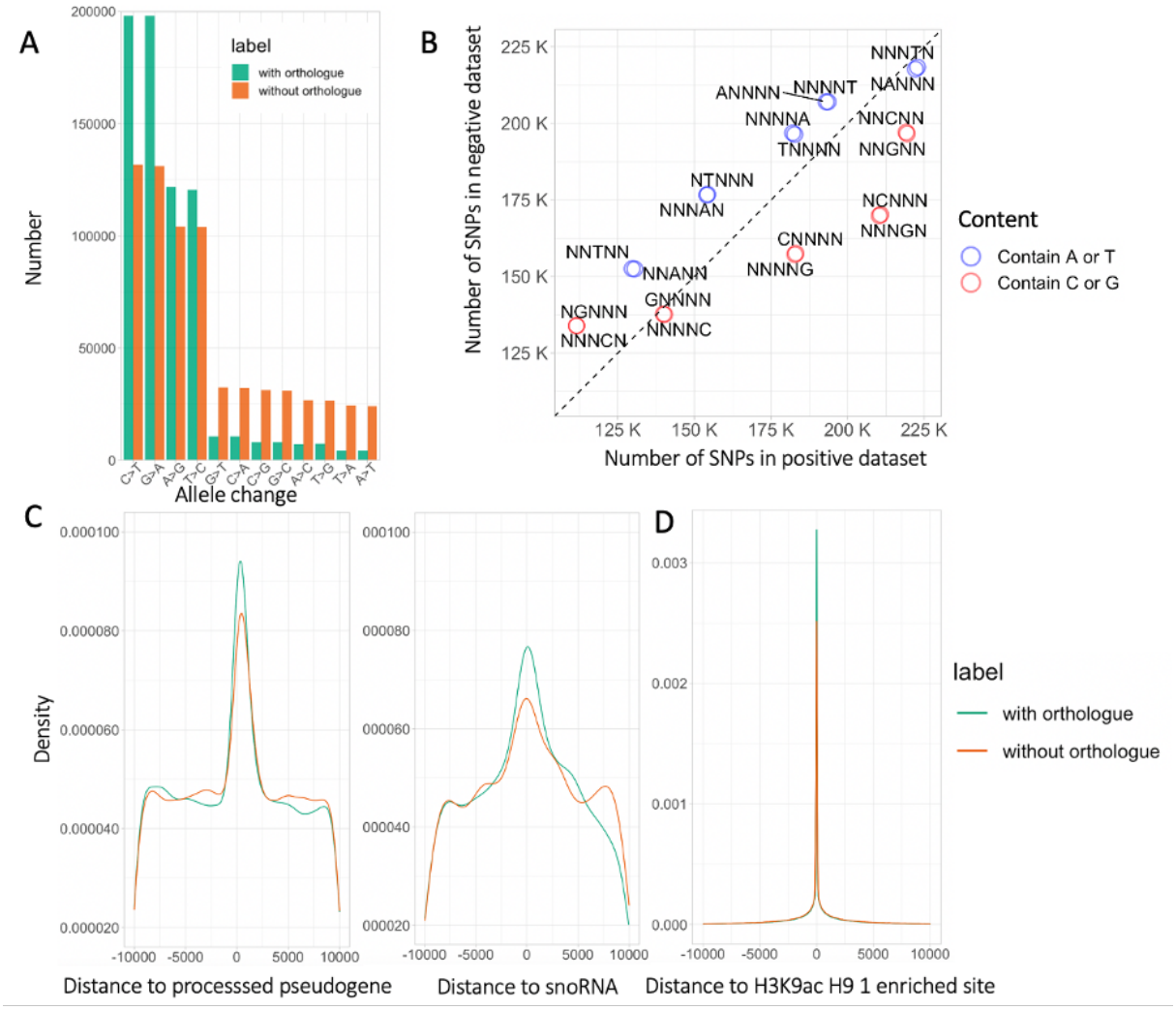


Figure 2. The characteristics of human variants with livestock orthologues. The genomic distribution of 1,397,362 human variants is shown (698,681 human variants with orthologues in cattle and an equally sized random sample of 698,681 human variants without orthologues in cattle). (A) Number of variants with or without orthologues by their observed allele changes (reference > alternative). (B) Number of SNPs with different 5-mer flanking sequences among variants with or without orthologues. Each circle represents a 5-mer flanking sequence with a specific base at a certain position, and the circle color indicates whether the specific base is C/G or A/T. The black dashed line represents parity, i.e., where the number of SNPs in the positive dataset equals the number of SNPs in the negative dataset. All the 5-mer sequences are significantly different between the groups at a P value less than $2.2x10^{-16}$ (Chi-Squared test). (C) Density plots of distances of variants with or without orthologues to processed pseudogenes and snoRNAs (plot restricted to within 10kb). Distances of variants to processed pseudogenes and snoRNAs are different between groups at P values less than $3.2x10^{-5}$ (Two-sample Kolmogorov-Smirnov test). (D) Density plot of distance between variants with or without orthologues to chromatin regions marked by H3K9ac in the human H9 cell line (plot restricted to within 10kb). Distance to these regions is different between groups at P value less than $1.8x10^{-3}$ (Two-sample Kolmogorov-Smirnov test).

Modelling the distribution of shared variants across the genome

Using 1589 different annotations (Table 1), including sequence conservation, chromatin context, and the distance to genome features such as genes, we compared the genomic distribution of human variants that do and do not have livestock orthologues in these cohorts. Several factors were observed to be associated with the probability of a human SNP having a livestock orthologue, including their distance to known genes and chromatin, and sequence context. For example, C to T changes are enriched among the variants with orthologues, with the hypermutability of CpG sites increasing the chance of the same change occurring across lineages (Figure 2A). More generally, human changes with a G:C base pair within their 5-mer flanking sequence are more likely to have a cattle orthologue than those with an A:T base pair at the same position (Figure 2B). A notable exception to this is where a guanine is found 5 prime of the human SNP site, with such changes less likely to have an orthologous SNP at the same position in cattle (Figure 2B). Variants with orthologues are also more likely to be enriched near specific genes such as processed pseudogenes and snoRNA, and around certain chromatin marks (Figures 2C, D).

We investigated the extent to which it is possible to use these genomic annotations to predict whether a human variant will have an orthologue in a livestock species. To do this, we used 140,000 human variants with or without a cattle orthologue and trained three tree-based machine learning models (Random forest, XGBoost and CatBoost, see methods) on the 1589-human genomic features (Table 1). To compare the performance of these models at discriminating human variants with and without cattle orthologues we tested the models on a further 60,000 human variants with the same

173 split that had not been included in the original training data. As shown in Figure 3A,
174 CatBoost outperformed the other models with an area under the receiver operating
175 characteristics (AUC) score of 0.69, an accuracy of 0.64 and an F1 score of 0.70.
176 These results suggest that the genomic annotations of the variants contain
177 discriminating information that makes it possible to identify which human variants have
178 a higher probability of having an orthologue in another species.

179 Models trained on human variants with orthologues in other species, such as pig, could
180 predict the presence of orthologues in these other species with similar accuracies as
181 the cattle specific models (Figure 3B+C). Likewise models trained on human variants
182 with orthologues in given species were largely as accurate at predicting orthologues
183 in completely different species (Figure 3D). This suggests the features associated with
184 orthologous variants are fundamental across mammals.

185 Comparison of the top 30 most important features of the three different modelling
186 approaches shown in Figure 3A found that the allele change, 5-mer flanking sequence
187 and conservation score (phyloP100way) were consistently three important features
188 (Figure 3E). Figure 3F shows the top 30 most important features of the cross-species
189 model, i.e. trained using human variants with an orthologue in any of the tested
190 livestock species, and how their values affect the predictions of the model. Sequence
191 conservation is the most important variable, with human variants in less-conserved
192 regions more likely to have an orthologue in another species. This is consistent with
193 mutations in these regions less often being removed, increasing the probability of the
194 same change occurring in different mammalian lineages. As well as the type of base
195 change the flanking sequence disproportionately contributes to the model

performance, with a G base at the position immediately downstream of the polymorphic site (NNNGN) associated with an increased probability of the same change being observed in the other species, consistent with the preferential deamination of CpG sites.

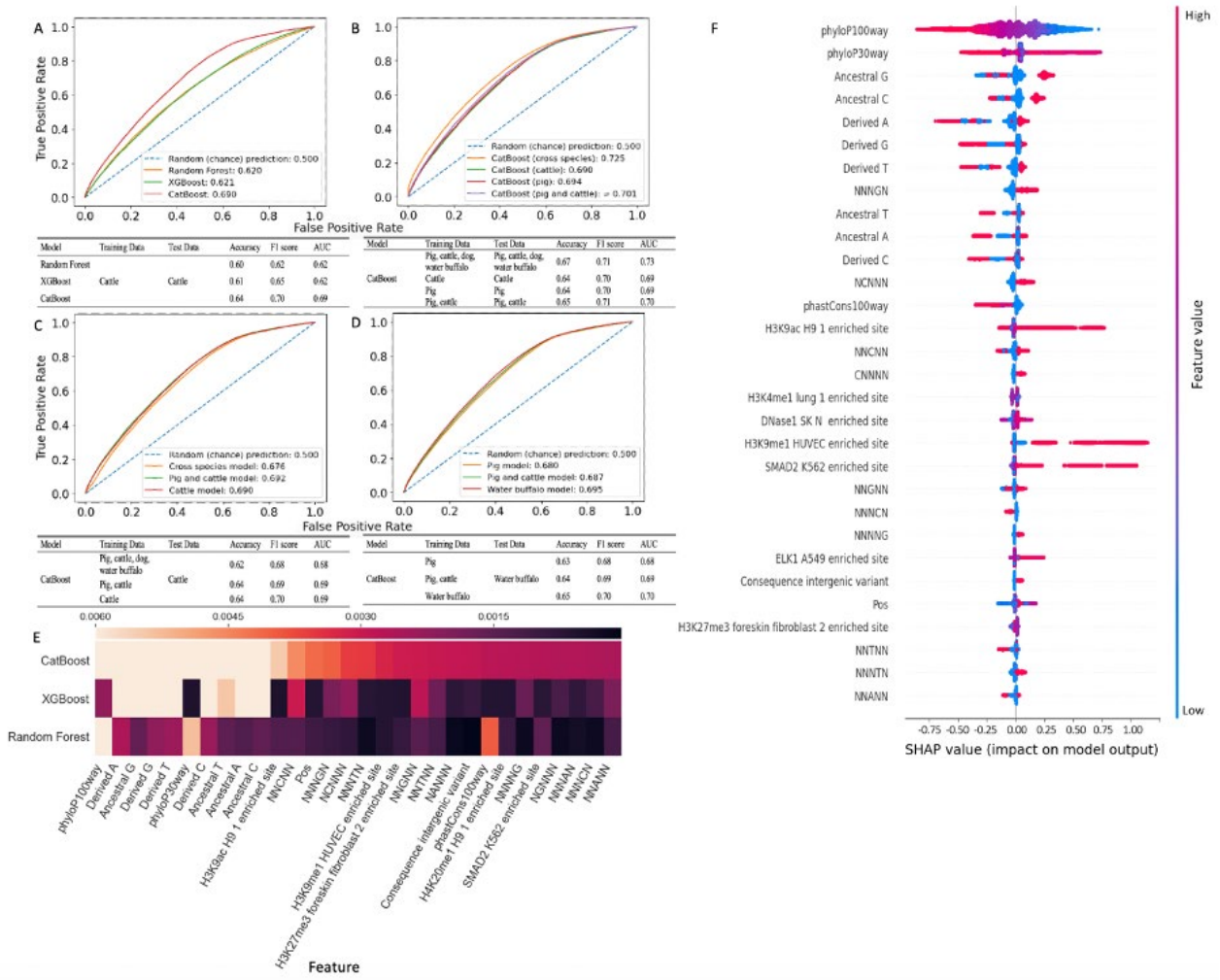


Figure 3. Machine learning models of orthologous variants. (A) Receiver operating characteristic (ROC) curves of Random Forest, XGBoost and CatBoost models trained and tested using human variants with and without orthologues in cattle. The numbers in the legend are area under the receiver operating characteristics (AUC) scores of the different models. AUC reflects a model's general ability of distinguishing between the classes. The table below the plot is the summary statistics of the experiment. (B) ROC curves and summary statistics of CatBoost models trained and tested using human variants with and without orthologues in cattle; pig; pig or cattle; pig, cattle, dog or water buffalo (cross species). (C) ROC curves and summary statistics of CatBoost models trained using human variants with and without orthologues in cattle; pig or cattle; pig, cattle, dog or water buffalo, but tested using human variants with and without orthologues in cattle. (D) ROC curves and summary statistics of CatBoost

models trained using human variants with and without orthologues in pig; water buffalo; pig or cattle, and tested using human variants with and without orthologues in water buffalo. (E) Feature heatmap of CatBoost, XGBoost, Random Forest models trained and tested using human variants with and without orthologues in cattle. Thirty important features are included in the figure, with lighter color indicating greater importance. (F) SHAP summary plot (Lundberg & Lee, 2017) of the CatBoost model trained using human variants with and without orthologues found in any of cattle, pig, dog and water buffalo. Features are ranked in descending order according to their importance on the left. The color represents low (in blue) and high (in red) value of the feature and the effect of their values on the output of the model is reflected by their positions on the x-axis.

Animal models of human pathogenic variants

Consequently, over a million human variants have a livestock orthologue. The modelling results highlight that these disproportionately fall in less conserved genomic regions, raising the question as to how many naturally occurring livestock models of human pathogenic variants exist. To characterize specifically how many human pathogenic mutations are segregating in other livestock species we first extracted 70,083 SNPs from the human Clinvar (Landrum et al., 2018) database labelled as “pathogenic” or “likely pathogenic”. Being mostly found in conserved coding regions, the overwhelming majority (99.4% and 94.3%) of these variants could be successfully mapped to an orthologous position in the Cow (BosTau9) and Pig (SusScr3) genomes. Using the data from the same cow and pig cohorts we identified how often these variants overlapped an orthologous variant in one of these other species. In total 1126 Clinvar variants overlapped a variant in the cow dataset and 673 in the pig cohort, of which 222 and 181 respectively also showed the exact same allele change observed in humans. In agreement with the modelling results, these numbers differ from those expected from the background numbers. Not only is the number of Clinvar variants with an orthologue in one of these livestock species substantially lower than expected

given the number of all human variants with an orthologue in pig or cattle, but also where a variant does segregate at the orthologous position it is less likely to show the same allele change. In total Clinvar variants are approximately three times less likely to have a variant at the orthologous position in either pig or cattle than expected from the frequencies in the 1000 genomes cohort, and approximately seven times less likely of having one displaying the same allele change (Supplementary Figure 2). These

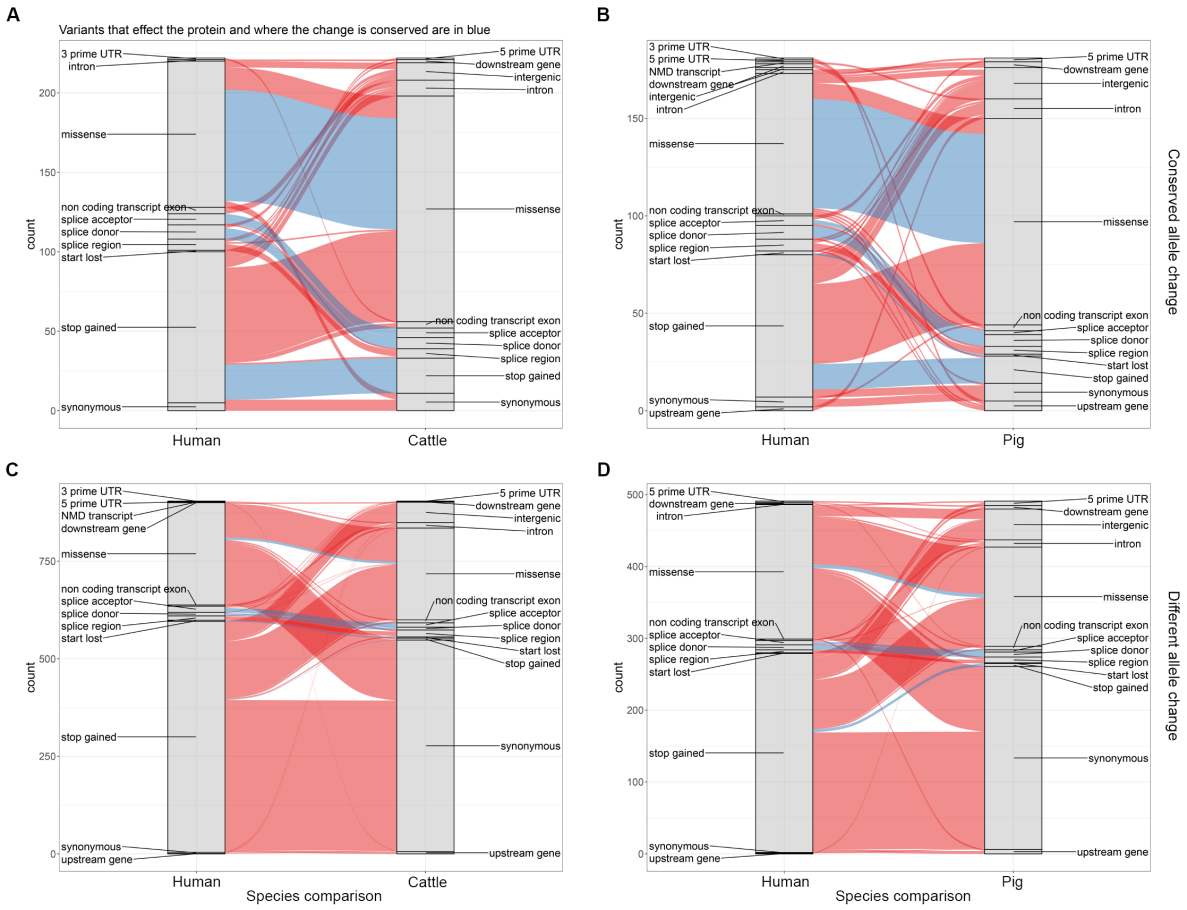


Figure 4. Conservation of impacts of orthologues of Clinvar variants. (A) The conservation of impact of genic orthologous variants across human and cattle where variants show the same allele change. The width of each ribbon indicates the number of variants with the given combination of consequences across the two species. Orthologous pairs of variants with conserved impacts on a protein are coloured blue. (B) The same as A but for human-pig orthologous variants. (C) Conservation of impact of variants across human and cattle where their locations are orthologous but they show different allele changes. (D) The same as C but for human-pig orthologous variants. The underlying data for all plots are provided in Supplementary Table 1.

241 results are consistent with these changes being deleterious as indicated, and selection
242 preferentially removing them across species.
243
244 Orthologous variants, even with the same allele change, may not have the same
245 impact on genes, if for example the gene structure and codons have changed between
246 species. As shown in Figure 4A, 76% of the 92 cattle orthologues of human Clinvar
247 variants leading to a missense change show the same missense change across both
248 species. A further 15% are missense in both species but involving different amino acid
249 changes. Only 4.3% of the human missense variants are predicted to be synonymous
250 in cattle, suggesting the consequence of human missense changes is most often
251 conserved across these mammals, with similar patterns observed in pigs (Figure 4B).
252 However, of 95 human Clinvar variants predicted to lead to the introduction of a stop
253 codon, only 23% also lead to a stop gained change in cattle, with the majority (60%)
254 predicted to just lead to an amino acid change due to a difference in the codon
255 between species. This may represent a true difference in the impact of these variants
256 between species, but may also sometimes reflect the comparatively poor annotation
257 of gene isoforms in livestock species. Of note, 27 cattle and 19 pig variants lead to the
258 same protein impact as their orthologue in humans despite involving a different allele
259 change (Figures 4C and D). Consequently, although rare, variants do not necessarily
260 need to show the same allele change to have a conserved impact.

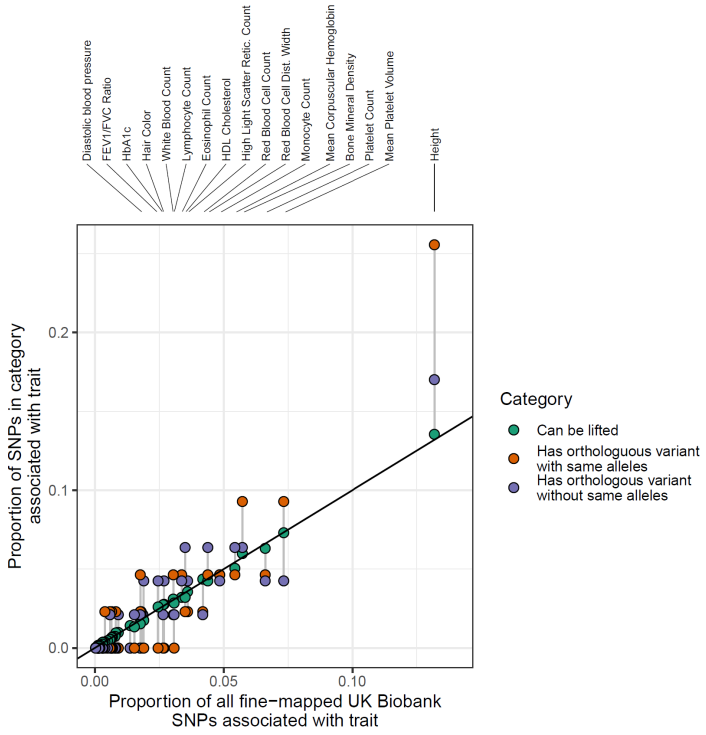
261 These data highlight that there are existing animal models available for at least several
262 hundred human Clinvar variants, including those linked to a variety of important
263 phenotypes such as cancers and Parkinson's disease (Supplementary Table 2).
264 Interestingly, Clinvar variants linked to certain traits are more likely to be found across

265 species. This includes those linked to biotinidase deficiency (Chi-squared test P <
266 $1x10^{-7}$), neurofibromatosis ($P=1.4x10^{-6}$) and glycogen storage in cattle (P=0.0004) and
267 factor VII deficiency in pigs ($P < 1x10^{-7}$). Of 23 known human biotinidase deficiency
268 variants, four (17%) have a direct orthologue showing the same allele change in cattle.
269 This is despite only 1.5% of all lifted Clinvar variants having a cattle orthologue. All
270 four of these variants are missense SNPs showing the same amino acid change in
271 both species, with one of the mutations having risen to a minor allele frequency of 22%
272 in cattle despite being found at a frequency of only 0.002% in humans, meaning
273 studying its impact may be easier in cattle populations. Supplemental biotin is often
274 fed to cattle as it is thought to improve hoof health and increase milk production (Lean
275 & Rabiee, 2011) and these variants are consequently also strong candidate functional
276 variants for further investigations for improvement of these cattle traits.

277 ***Animal models of common variants of polygenic diseases and traits***

278 We next investigated whether there are potential existing livestock models of common
279 human variants linked to polygenic diseases and traits. To do this, we obtained 2240
280 fine-mapped SNPs linked to 47 different traits in the UK biobank cohort (Weissbrod et
281 al., 2020). In total 58 of these variants had a direct orthologue in either pigs or cattle.
282 Interestingly variants linked to height in humans were significantly more likely to have
283 a direct orthologue in cattle than other traits, with over a quarter of variants (11 out of
284 43) that were found in both species with the same alleles being linked to this phenotype
285 (Figure 5, Supplementary Table 3). This is compared to only 13.6% (341 out of 2513)
286 of the variants successfully mapped between the species being linked to this trait (two-
287 tailed Fisher's exact test P=0.040). Of these 11 variants, 3 are missense changes

288 (rs154001, rs61735104, rs79485039), with each leading to the same amino acid
289 change in both species. These amino acid changes fall in FGFR3, KIAA1614 and

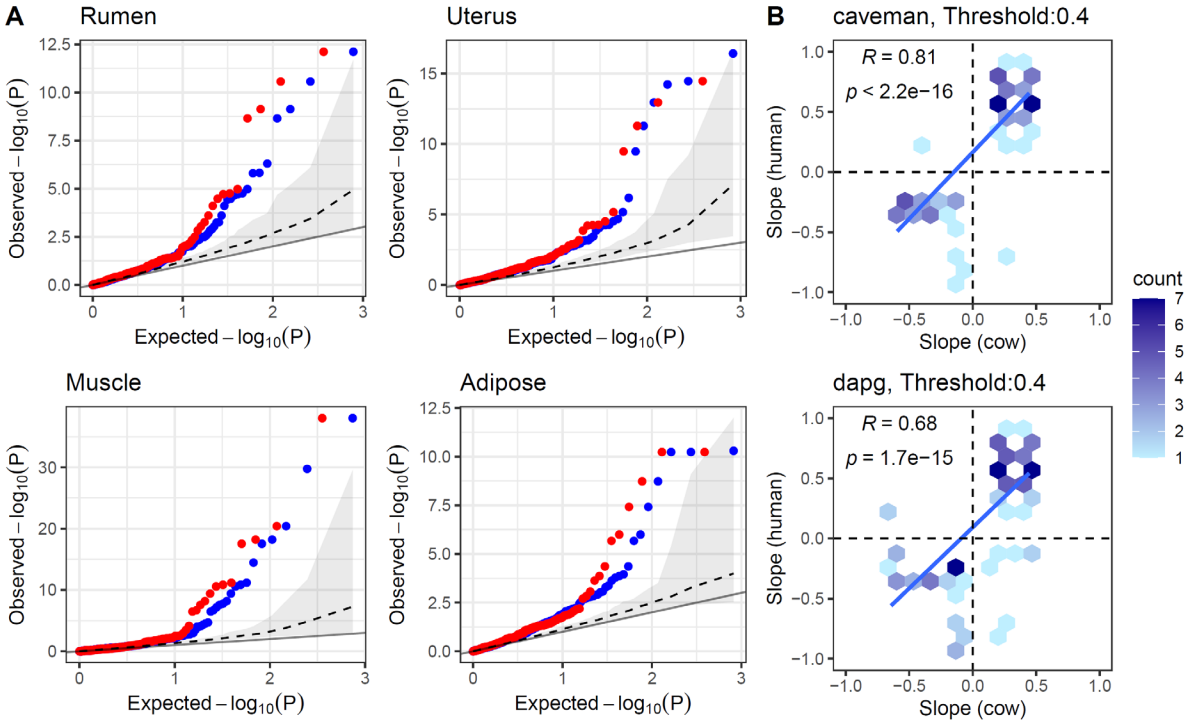


290

291 **Figure 5. Orthologues of fine-mapped variants linked to traits in the UK Biobank.** All 2240 fine-mapped UK
292 biobank variants were lifted over to the cattle genome and the number that overlapped a variant in the cattle
293 genome with and without the same alleles were determined. The proportion of all lifted variants that were
294 associated with a given trait is strongly correlated with the original proportion of fine-mapped variants linked to that
295 trait (green circles). However, variants with an orthologous cow variant are disproportionately associated with
296 height, and in particular those with matching alleles in both species (orange circles). Circles corresponding to the
297 same trait are connected by vertical grey segments, with the trait indicated above for those traits with at least 55
298 fine-mapped variants.

299 FBN2, with a further gene, *FOXM1*, having a variant (rs28990715) at orthologous
300 positions in both species that leads to the same amino acid change despite having
301 different allele changes. 10 of the 11 human variants with cattle orthologues were in
302 the Gene Atlas UK Biobank (Canela-Xandri et al., 2018) results and are together,
303 under certain assumptions, associated with a predicted 2.7cm variation in human
304 height. This corresponds to around ~1/3 of a standard deviation of the human heights
305 in the UK Biobank cohort. The FGFR3 change alone is associated with a ~1cm

306 difference in standing height between opposing homozygotes. Mutations in FGFR3
307 underlie 99% of cases of human achondroplasia that affects bone development and
308 leads to short stature, but the role of this gene in cattle stature is less well
309 characterized (Wilkin et al., 1998). Potentially, in part, because all 11 variants are rare
310 in cattle (10 with a frequency of less than 1%, with 1 with a frequency of 6%) and would
311 be unlikely to be detected in a standard cattle GWAS, but these variants are
312 consequently strong candidate functional rare variants for contributing to variability in
313 cattle height due to their strong associations with this trait in humans that could be
314 exploited to alter cattle stature.



315

316 **Figure 6. Conserved effects of regulatory variants across humans and livestock.** (A) Quantile-quantile (Q-Q)
317 plots of observed and expected cis-eQTL P values of cattle variants that are direct orthologues of human fine-
318 mapped regulatory variants. The blue points represent the observed and expected P values of the cattle variant's
319 association with the expression level of the cattle orthologue of the corresponding human gene in four cattle tissues
320 irrespective of its allele change. The red points are the same after restricting to these variants exhibiting the same
321 allele change as observed at the human SNP. The grey dashed line and grey ribbon represents the median and
322 95% confidence interval obtained when randomly sampling the same number of variants as shown by the blue
323 points from all cattle variants tested (irrespective if have a known human orthologous variant or not) 1000 times.

The line of parity (solid black) is also shown. This illustrates cattle orthologues of human fine-mapped regulatory variants are more likely to show evidence of also being linked to the orthologous gene's expression across different tissues. (B) Comparison of the slopes (direction of effects) of eVariants across species. The slopes of human fine-mapped regulatory variants (using caveman approach, top, or dap-g, bottom) were compared to the slopes observed for their orthologues if they also had a significant cattle GTEx association with the expression level of the orthologous gene in cattle. The slope represents the impact on the gene's expression of increasing the dosage of the same allele in both species. Note the same eVariant can be found in multiple tissues and can therefore be represented multiple times in this plot. In total there are 77 and 106 human-cattle association pairs in the caveman and dapg plots, involving 36 and 57 distinct human eVariant-gene-tissue associations. The significant positive correlation remains if only one entry for each human eVariant-gene-tissue association is retained. This agreement in direction is seen despite not restricting to comparing the effects to the same tissues across species, i.e. the direction of effect is generally conserved across tissues as well as species.

Conservation of regulatory variation

Most variants linked to important complex phenotypes are thought to be regulatory rather than coding (Cano-Gamez & Trynka, 2020). To investigate whether regulatory variants are conserved across species we obtained the location of fine-mapped regulatory SNPs from the human GTEx (GTEx Consortium, 2020) dataset. These human regulatory variants had been fine mapped using three different approaches; CAVIAR (Hormozdiari et al., 2014), CaVEMaN (Brown et al., 2017) and DAP-G (Wen et al., 2017), and we took the superset of SNPs across all three. We then extracted the associations with orthologous genes of variants found at the orthologous location in cattle from the cattleGTEx project, who defined eQTLs across 23 different cattle tissues and cell types (Liu et al., 2021). In total 185 of the human-fine mapped variants had a matching cattle variant in the cattleGTEx data that had been tested against the same orthologous gene in at least one tissue. Ignoring the allele change of the variants this number increases to 391. As shown in Figure 6A, these cattle variants at the

orthologous position of the human fine-mapped variants are more likely to show an association (i.e. have a smaller P value) than randomly sampled gene-variant pairs from the cattleGTEx cohort. This suggests these variants are often regulatory across both species. Notably this was largely observed whether restricting the cattle variants to those showing the same allele change as the human variant or not, with only a slight enrichment of smaller P values among the former group in some tissues (Figure 6A). This suggests simply disrupting the same regulatory site may often be sufficient to affect the gene's expression across species in many cases.

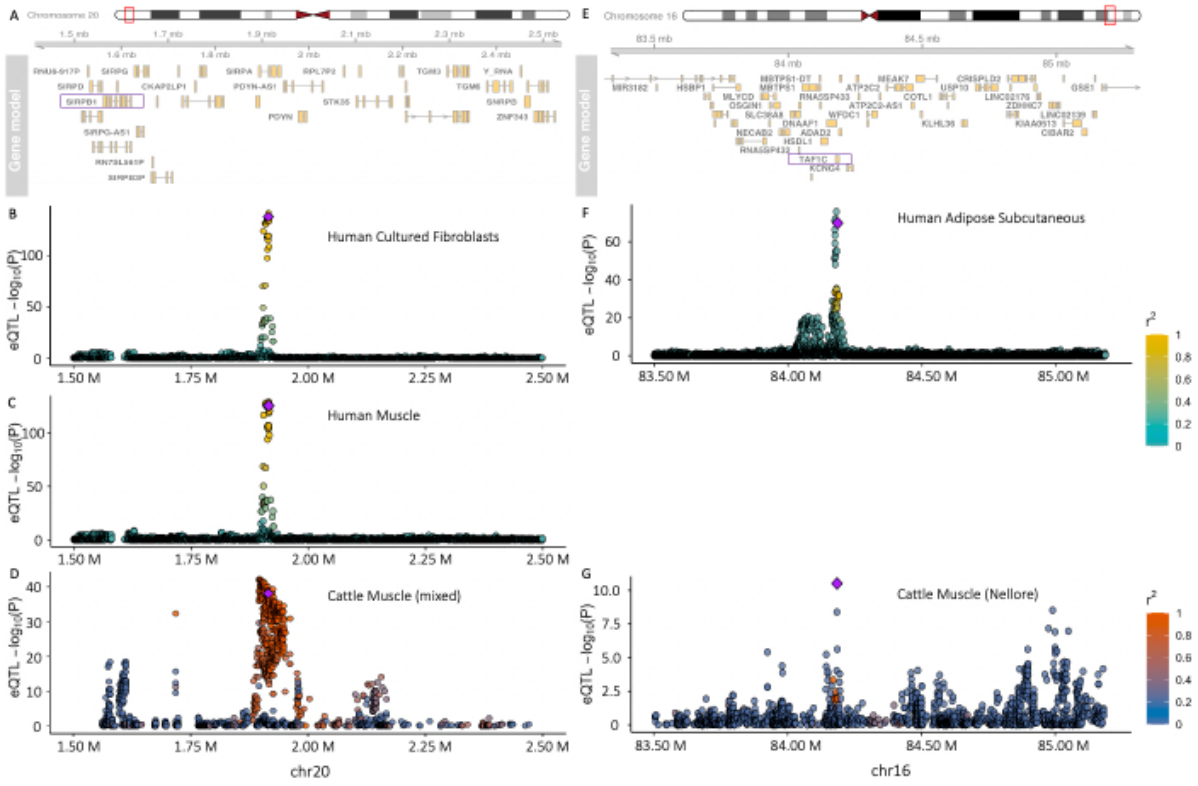


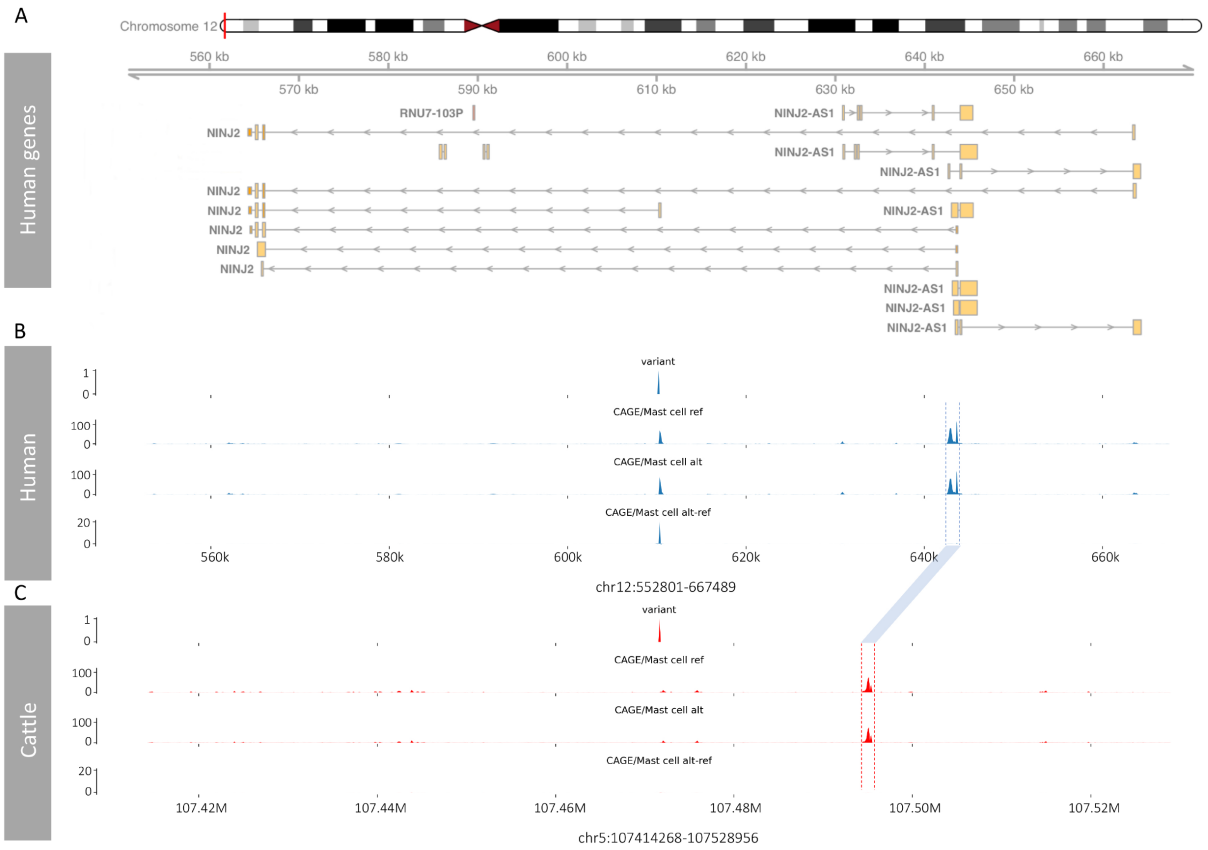
Figure 7. Colocalization of eQTLs across humans and cattle. (A) The gene neighborhood of a shared eQTL rs115287948 across both humans and cattle. The gene regulated by the eQTL is indicated by a purple rectangle. (B) Strength of association of human variants with *SIRPB1* expression levels in cultured fibroblasts tissue. The fine-mapped regulatory variant, rs115287948, with a cattle orthologue is represented by the purple diamond. Other variants are colored according to their linkage disequilibrium (r^2) with this variant. (C) Strength of association of the same human variants but in muscle tissue. (D) Strength of association of variants with *SIRPB1* in cattle muscle tissue (mixed breeds). Each variant is plotted according to their orthologous position in the human genome and the

variant with a fine-mapped orthologue in the human GTEx data is represented by the purple diamond. (E), (F), (G), same as (A), (B), (D) but for variant rs2230126 linked to the expression of *TAF1C* in different tissues.

The direction of effect of conserved regulatory variants were more likely to be conserved across the two species (Figure 6B), i.e. the same alleles are associated with increased or decreased expression across species. This confirms that the effect of predicted functional variants appear often conserved. This is despite the different linkage disequilibrium patterns between the species, that may be expected to disrupt any conservation of direction of effects if these variants were not functional.

Figure 7 shows examples of the colocalization of eQTLs across humans and cattle. rs115287948 is a missense variant in the *SIRPA* gene that was finemapped in the GTEx cohort as a causative regulatory variant (probability > 0.5) linked to the expression of *SIRPB1* across a range of tissues including cultured fibroblasts and muscle (Figure 7A,B,C). A direct orthologue of this variant is also found at a co-localised eQTL in cattle muscle (Figure 7D) displaying an association with the same gene with the same direction of effect.

rs2230126 is a variant falling within an alternative promoter of *TAF1C* with which it is a fine-mapped human regulatory variant (probability > 0.95) in a range of tissues including subcutaneous adipose (Figure 7E,F). An orthologue of this variant is also found in cattle and is the lead eVariant for the same gene in Nellore muscle tissue (Figure 7G).



386 **Figure 8. Enformer predictions for human and cattle.** (A) The human gene neighborhood of variant rs10849334.
387 Different isoforms of the genes are included in the plot. (B) Enformer predicted Cap Analysis Gene Expression
388 (CAGE) tracks of variant rs10849334 in mast cells (the track showing the largest human alt-ref difference). The top
389 track shows the position of the target variant in the human genome. The two tracks in the middle are the predicted
390 CAGE levels for the reference and alternative alleles from Enformer when run on the human sequence. The bottom
391 track shows the predicted difference between the CAGE levels from the reference and alternate sequences
392 specifically at the shorter *NINJ2* isoform. (C) Predicted CAGE tracks derived from the cattle sequences around the
393 orthologous variant of rs10849334. The orthologous peaks at the TSSs of the longer *NINJ2* transcripts in cattle
394 and human are indicated by the blue linking bar. No CAGE peak is predicted at the TSS of the shorter isoform.

395 We explored why some variants conserved in both species may not show evidence of
396 impacting gene expression in cattle. Figure 8 illustrates predictions from the Enformer
397 human deep learning model (Avsec et al., 2021), that predicts transcriptional potential
398 and chromatin states from DNA sequence alone. As shown in Figure 8A,B, Enformer
399 predicts that the alternate allele of the rs10849334 variant is associated with reduced
400 expression, specifically of an alternate, shorter isoform of the *NINJ2* gene. The

401 predicted transcriptional potential of the TSS of longer isoforms of the *NINJ2* gene are
402 unaffected by this variant. When the orthologous cattle DNA segment is run through
403 Enformer the predicted transcriptional potential of the TSS at these longer isoforms
404 remains, but the CAGE peak at the promoter of the shorter isoform is completely
405 abrogated. Consistent with this we could also not find any evidence of a cattle
406 orthologue of this shorter isoform in the public databases. Consequently the lack of
407 evidence of the cattle orthologue of rs10849334 affecting the expression of *NINJ2* is
408 potentially due to the human variant specifically regulating this shorter isoform, that is
409 absent in cattle due to sequence divergence in this locus.

410 In summary, although the effect of some variants do not lift over across species,
411 potentially due to for example changes in the usage of isoforms between species, a
412 range of human regulatory variants have orthologues in cattle that often have
413 conserved effects and can consequently be used to provide insights into their
414 mechanisms of gene regulation.

415 **Discussion**

416 In this study we have demonstrated how millions of orthologues of human variants
417 exist in domesticated species, including hundreds of orthologues of fine-mapped
418 functional variants linked to diseases and phenotypes. These are consequently readily
419 accessible large animal models of important human variants, that can potentially be
420 studied at scale without the time and costs associated with transgenic approaches.
421 Importantly we show that orthologous regulatory variants most often have conserved

422 directions of effect across humans and cattle, suggesting their downstream effects can
423 be effectively studied in these species.

424 Variants shared across humans and domesticated species are not restricted to one
425 type of trait, but are linked to a wide spectrum of phenotypes. From rare, monogenic
426 disorders such as cystic fibrosis to highly polygenic traits such as height. However,
427 variants associated with particular phenotypes are found across species more often
428 than expected. For example, the 4 out of 23 variants associated with biotinidase
429 deficiency that have a direct orthologue in cattle. It is in unlikely the co-occurrence of
430 these variants is purely due to, for example, a higher mutation rate around the gene
431 linked to this phenotype, as none of the other three species studied carry even one
432 orthologue of these variants. This suggests there is a preferential overlap of variants
433 linked to specific phenotypes, including more polygenic phenotypes such as height.
434 These likely reflect selection to preferentially maintain such variants. This not only
435 provides insights into the evolution of these species, but also potential candidate
436 variants for livestock breeding programs. The increased number of human height
437 associated variants with an orthologue in cattle likely reflects the selection for body
438 size in domesticated animals. However, as these variants remain polymorphic, the
439 selective sweep is incomplete, and they remain suitable targets for breeding programs.
440 Although there would be potential ethical concerns of introducing human variants into
441 livestock species to improve their production, the same is not true if the variant already
442 exists in the species, as even editing the variant into another breed, would no longer
443 come with the restrictions imposed on transgenic projects. Consequently, exploiting
444 the large amount of data and studies on functional variants in humans, could
445 potentially be leveraged to prioritise variants for testing their effects in livestock.

Despite the hundreds of orthologues of functional variants identified, this number is likely a substantial under-estimate of the true number of shared functional variants. This is because for a variant to be tied to a phenotype it needs to not only be at a sufficiently high frequency in the population to be discovered, but also with suitable data and patterns of linkage disequilibrium to be fine-mapped. However, most functional variants are, by their nature one or more of: rare, non-coding or in regions of elevated LD, and are therefore difficult to tie to a trait. Looking at the effects of rare variants across species may though help increase the pool of individuals in which to study their potential role. Likewise looking across species has the advantage that allele frequencies and linkage disequilibrium patterns can differ substantially and may, therefore, help in fine-mapping approaches. Extending this further, such approaches may help validate fine-mapping methods, for example by characterising which fine-mapping approach better identifies variants whose impacts are subsequently shown to be conserved across species. This is illustrated by the comparison of fine-mapped regulatory variants, and their conserved direction of effect across species. Of the three fine-mapping approaches studied, CAVIAR fine-mapped variants showed the lowest conservation of effect direction. This may reflect where variants are not truly functional, with the different patterns of LD in the different species with the actual causative variant meaning their eQTL coefficients are less conserved.

A caveat to such cross-species comparisons of regulatory variants is that not only can it be difficult to directly match tissues between species, but that power also generally differs due to differences in sample sizes. Consequently, the fact that a variant doesn't also show evidence of being linked to a gene's expression in another species doesn't mean it is not a functional variant in both. Those variants we detect as being linked to

gene regulation across species are likely those that are regulatory in multiple tissues, increasing the probability of us detecting it's association in at least one.

It is likely that few if any human polymorphisms with orthologues in these mammalian species arose prior to the divergence of the respective species, and to still be polymorphic down the independent lineages. Rather they will have largely arisen independently in each. This is supported by the fact that despite there being millions of orthologues of human variants segregating in other mammals, few are found in more than one other species. Only twelve sites were polymorphic across the five studied mammals. Shared variants are simply most often found at sites with the highest mutation rates and lowest levels of purifying selection, and therefore reflect the increased chance of these sites mutating and not being purged from the population in both species. This indicates that the normalised presence or absence of orthologous variants may provide an alternate metric of the selective pressure on genomic regions, as illustrated by the depletion of orthologues of Clinvar variants across species.

Consequently the study of orthologues of human functional variants can be used across a range of studies. From understanding the biological mechanisms linking variants to important downstream phenotypes, to providing potential targets for livestock breeding and genome editing programs as well as understanding the selection pressures on our species.

Methods

Genotype Datasets

Previously published and filtered genotype data for five different species was used in this study. The genome-wide set of 78 million human SNPs from 2,504 individuals was obtained from the 1000 genomes consortium (The 1000 Genomes Project Consortium et al., 2015). The dog genotypes from 722 individuals were obtained from Plassais et al (Plassais et al., 2019). The cattle and water buffalo genotypes of 477 and 79 individuals, respectively, were obtained from Dutta et al (Dutta et al., 2020), and the pig genotypes from across 409 individuals from the Genome Variation Map website (C. Li et al., 2021). All cohorts were subsequently restricted to biallelic SNPs only (cattle: 88,067,584; pig: 90,901,469; water buffalo: 37,682,631; dog: 73,906,017). For all sets of human variants, their positions were lifted to their orthologous positions in the pig (SusScr3), cattle (BosTau9) and dog (CanFam3) genomes using the UCSC liftover utility (Hinrichs et al., 2006) with chain files available from the UCSC website. For the water buffalo (Low et al., 2019), where no public chain file exists, we used the nf-LO pipeline (Talenti & Prendergast, 2021) to perform the liftover. Sites that were lifted to more than one location were excluded. SNPs from other species were said to have the same allele change as human SNPs if found at the orthologous position with alleles that directly matched or that matched their complement. This therefore assumes the ancestral base in these conserved regions is the same across mammals. To test the impact of relatedness on the number of orthologous variants found in cattle, we used the relatedness2 (Manichaikul et al., 2010) parameter in vcftools (Danecek et al., 2011) to identify pairs of animals with a kinship coefficient greater than 0. Individuals in each pair were then iteratively removed till the kinship coefficient between all pairs of remaining animals was 0 or less.

Clinvar and UK biobank analyses

The location of variants potentially linked to human health were downloaded from Clinvar (Landrum & Kattman, 2018), which contains SNPs linked to different human clinical phenotypes. Restricting this set to those labelled as “pathogenic” or “likely pathogenic” left 76,752 SNPs with likely functional consequences. Potentially functional variants linked to polygenic traits were obtained from Weissbrod et al. (Weissbrod et al., 2020). This study produced a list of 3281 fine-mapped, potentially functional variants associated with 47 complex traits of which 2240 were SNPs. The locations of these sets of SNPs were intersected with those from other species to identify those segregating in other mammals as described above. To test whether Clinvar variants linked to particular phenotypes are more likely to segregate in another species than expected, we used a Chi-squared test to examine whether the proportion of successfully lifted Clinvar variants linked to a particular phenotype that overlapped a variant with matching alleles was significantly higher than that observed across all other phenotypes. To test whether UK biobank variants lifted to the cattle genome were disproportionately associated with height a Fisher’s exact test was used, comparing the proportion of variants with an orthologous variant with the same alleles that were linked to human height versus the proportion of successfully lifted variants linked to the same trait. To examine the impact of these orthologous variants on genes in humans, pigs and cattle, the variants were annotated using the Ensembl REST API in R, recording just the most severe reported consequence in each case.

Regulatory variant analyses

The GTEx v8 fine-mapped results for CaVEMaN, DAP-G and CAVIAR were downloaded from the GTEx portal. Together these reported 5,341,519 distinct tissue-gene-variant associations of which 2,145,167 could be lifted to an orthologous position

539 in the cow genome. Upon filtering out variants that did not have a minimum probability
540 > 0.2 in at least one of the datasets, this number reduced to 230,991 associations.
541 These associations were then intersected with the cattleGTEx data to identify where
542 an orthologous variant was significantly associated with the corresponding
543 orthologous gene. Nominal P values were obtained as described in the original
544 cattleGTEx paper (Liu et al., 2021) and the human-cow gene orthologues were
545 obtained from Ensembl version 103 (Yates et al., 2020).

546 To examine whether cattle orthologues of human fine-mapped eVariants were more
547 likely to show evidence of also being significantly associated with the expression level
548 of the same gene, we extracted their corresponding P values from the cattleGTEx data
549 by tissue. The distribution of these P values were then compared to the distributions
550 of P values for the same tissue of the same number of variants sampled from the total
551 cattleGTEx data 1000 times to produce the shaded confidence intervals in the Q-Q
552 plots.

553 To conservatively estimate false discovery rates for the cattle eQTLs we used the
554 same random samples. For each real eQTL P value we divided the average number
555 of tissue-specific P values across the 1000 samples that had a P value as small or
556 smaller by the corresponding number within the variants that were orthologues of
557 human fine-mapped regulatory variants. This therefore corresponds to the
558 approximate probability of having sampled a P value as small or smaller from the
559 background list of all variants tested in the cattleGTEx project. This is conservative as
560 a large number of the variants in this background list are eVariants. Therefore, this
561 FDR corresponds to the false discovery rate above and beyond that expected given

562 the number of regulatory variants in the background, and variants with a large FDR
563 may still be regulatory variants.

564 To investigate why some regulatory variants shared across human and cattle may not
565 have conserved impact on gene expression in cattle, we used Enformer, a deep
566 learning architecture designed for predicting how DNA sequence influences gene
567 expression (Avsec et al., 2021). We loaded the trained Enformer model, made
568 predictions for reference and alternative alleles of each shared regulatory variant in
569 both human and cattle, and obtained 5,313 predicted genomic tracks for each variant.
570 The effect of each variant was evaluated by the difference between the reference and
571 alternative predictions.

572 ***Variant annotation and modelling***

573 The genome-wide set of 78 million human SNPs were annotated with 1589 features
574 across four categories (Table 1), including sequence conservation, variants position
575 properties, VEP (McLaren et al., 2016) annotations and sequence context. For
576 sequence conservation, we included 4 different conservation scores:
577 phastCons100way, phastCons30way (Siepel et al., 2005), phyloP100way and
578 phyloP30way (Pollard et al., 2010). We downloaded bigWig files of these conservation
579 scores from the UCSC genome annotation database (Navarro Gonzalez et al., 2021)
580 (hg38) and extracted the values at given positions using the pyBigWig python package
581 (Ramírez et al., 2016). To fully capture the position characteristics of the variants, we
582 calculated the distance between the variants and different genome elements. We
583 obtained the location of CpG islands from the UCSC genome annotation database
584 (Navarro Gonzalez et al., 2021), chromatin data (such as histone marks), TSS and

585 regulatory features (enhancer, promoter, CTCF binding site and TF binding site) from
586 Ensembl (Yates et al., 2020) (version 103). We used bedtools (Quinlan & Hall, 2010)
587 closest command to calculate distances to CpG islands and chromatin data. The
588 ChIPpeakAnno (L. J. Zhu et al., 2010) R package was used for getting distances to
589 the nearest TSS by biotype (only common biotypes were included, count >= 1000)
590 and distances to various regulatory features. Then we used the VEP (McLaren et al.,
591 2016) command line tool to annotate the variants and get the allele frequencies and
592 consequences of the variants. Instead of using Reference/Alternative alleles, we used
593 Ancestral/Derived alleles for the allele change. The human ancestral genome (hg38)
594 was downloaded from Ensembl (Yates et al., 2020) (version 103) and the bedtools
595 (Quinlan & Hall, 2010) getfasta command was used to extract the ancestral base. To
596 get the 5-mer flanking sequences centered on the target variants, we used the
597 samtools (H. Li et al., 2009) faidx command.

598 Using these genomic annotations as classification features, we trained machine
599 learning models to predict whether a human variant has an orthologue in other
600 livestock cohorts. Variants that have cattle orthologues (with matching alleles) were
601 used as the positive data in the models while variants without orthologues, i.e.,
602 variants that can be lifted to the cow genome but no cattle polymorphism was found,
603 were used as negative data. Similarly, we got positive and negative datasets for pig,
604 water buffalo, the intersection of variants found across the cattle and pig cohorts, and
605 the cross-species cohort (cattle, pig, dog and water buffalo). For the cross-species
606 cohorts, variants with orthologues in any of the tested livestock species were used as
607 positive data and variants without orthologues in all tested species were used as

608 negative data. To avoid class imbalance problems, we downsampled all negative
609 datasets to the same sizes as the positive datasets for all models.

610 The feature tables were pre-processed before being used for model training. Data with
611 missing values (found for sequence conservation scores) were discarded as they only
612 accounted for a small proportion (1.4%) of the whole dataset. Categorical features,
613 i.e., chromosome, consequence, allele change and 5-mer flanking sequences were
614 encoded using different encoding methods (see Table 1). To minimize the introduction
615 of new feature columns into the feature table and make the encoding more meaningful,
616 a self-defined binary encoding method was used for sequence context features. We
617 defined a dictionary for 4 bases (A: 1000, C: 0100, G: 0010, T: 0001) and mapped
618 each base in the sequences to the corresponding binary string. The final strings for
619 the sequences were split into binary columns and replaced the original categorical
620 features in the feature table. We constructed three tree-based machine learning
621 models, Random Forest (Breiman, 2001), XGBoost (Chen & Guestrin, 2016) and
622 CatBoost (Prokhorenkova et al., 2017) using the Scikit-learn (Pedregosa et al., 2011)
623 Python package. Models were trained on Eddie (*Edinburgh Compute and Data Facility*
624 *Web Site*, 2021), a compute cluster of the University of Edinburgh, and 2 64GB GPUs
625 on Eddie were used to train the CatBoost models. To enable balanced comparisons,
626 subsets (200,000 data in total, 100,000 of which was positive data and 100,000
627 negative data) of the datasets for different species were used. Each subset was
628 divided into a training set and test set at the ratio of 70% and 30%. We used 5-fold
629 cross-validation to evaluate our models on the training sets. To improve the
630 performance of the models, we used random search (Bergstra & Bengio, 2012) and
631 manual tuning methods for hyper-parameter tuning.

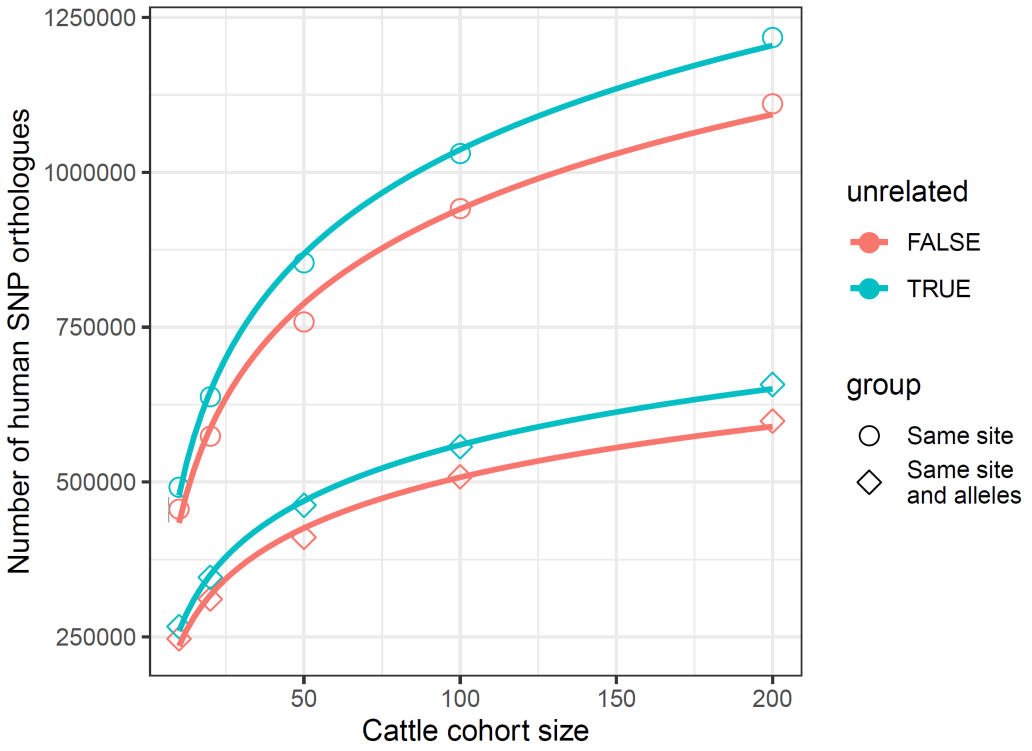
Table 1

Variant annotations and encoding methods

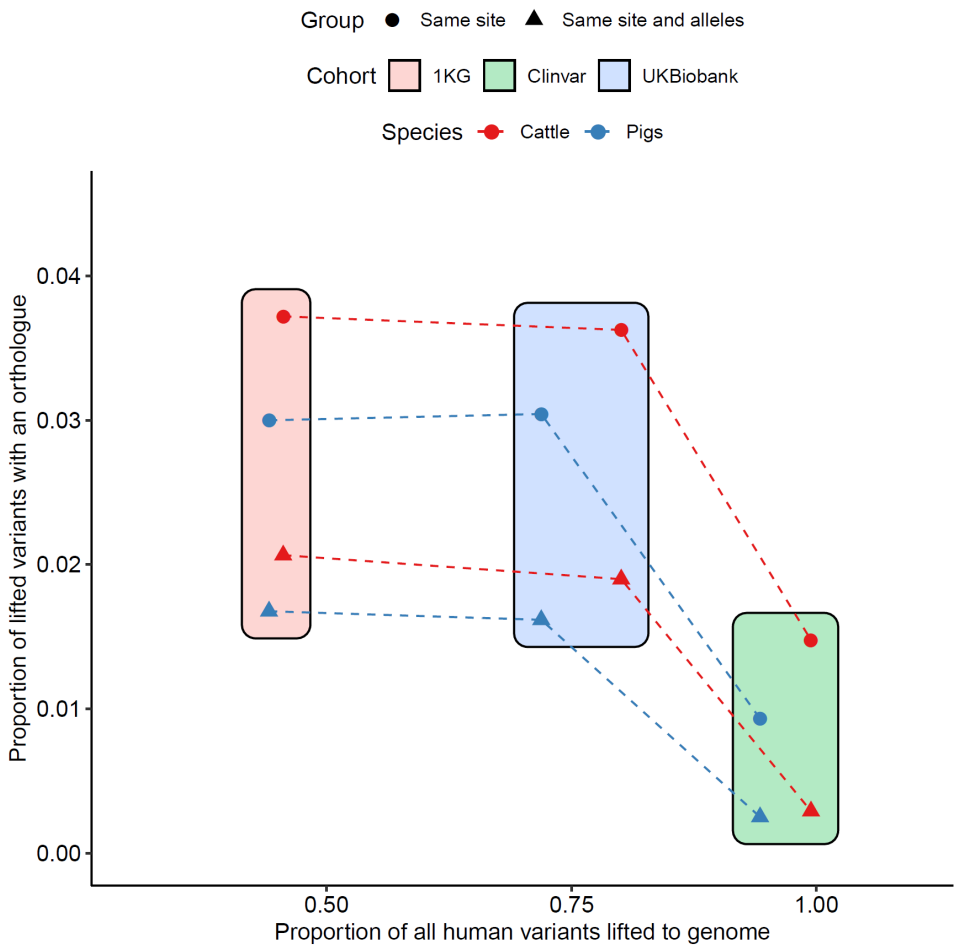
	Annotation	Encoding method	Number of features	Number of columns after encoding
Sequence conservation	phastCons100way	-	1	1
	phastCons30way	-	1	1
	phyloP100way	-	1	1
	phyloP30way	-	1	1
Variant position properties	Distance to CpG island	-	1	1
	Distance to chromatin data [a]	-	1554	1554
	Distance to TSS [b]	-	14	14
	Distance to regulatory features [c]	-	4	4
	Chromosome	One-hot encoding	1	22
	Variant position	-	1	1
	Gene density (per megabase)	-	1	1
VEP annotations	Consequence	One-hot encoding	1	34
	Allele frequency [d]	-	6	6
Sequence context	Allele change	Self-defined encoding	1	8
	5-mer flanking sequence	Self-defined encoding	1	20

[a] *Distance to 1554 different chromatin data from Ensembl.*
[b] *Distance to TSSs within 14 common biotypes (frequent >= 1000).*
[c] *Regulatory features include enhancer, promoter, CTCF binding site, TF binding site.*
[d] *A total of five allele frequencies from the 1000 genomes combined population and the African, American, East Asian, European and South Asian populations separately.*

Supplementary Figures



Supplementary Figure 1. The effect of cohort relatedness on the number of orthologues of human SNPs found in cattle. The red line shows the number of orthologues found when no filtering based on relatedness was applied to the cohort. The cyan lines show the effect of excluding from the cohort related animals so that all those remaining have a kinship coefficient (Manichaikul et al., 2010) with each other equal to or less than 0.



Supplementary Figure 2. The proportion of human variants that have a livestock orthologue by variant set. Although most Clinvar variants could be lifted to a position in the cow and pig genomes, fewer than expected had orthologues either with or without the same alleles. This likely reflects the strong selection against these changes.

References

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18(10), 1196–1203. https://doi.org/10.1038/s41592-021-01252-x

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. The Journal of Machine Learning Research, 13, 281–305.

Bouwman, A. C., Daetwyler, H. D., Chamberlain, A. J., Ponce, C. H., Sargolzaei, M., Schenkel, F. S., Sahana, G., Govignon-Gion, A., Boitard, S., Dolezal, M., Pausch, H., Brøndum, R. F., Bowman, P. J., Thomsen, B., Guldbrandtsen, B., Lund, M. S., Servin, B., Garrick, D. J., Reecy, J., … Hayes, B. J. (2018). Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nature Genetics, 50(3), 362–367. https://doi.org/10.1038/s41588-018-0056-5

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

Broekema, R. V., Bakker, O. B., & Jonkers, I. H. (2020). A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. Open Biology, 10(1), 190221. https://doi.org/10.1098/rsob.190221

Brown, A. A., Viñuela, A., Delaneau, O., Spector, T. D., Small, K. S., & Dermitzakis, E. T. (2017). Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. Nature Genetics, 49(12), 1747–1751. https://doi.org/10.1038/ng.3979

Canela-Xandri, O., Rawlik, K., & Tenesa, A. (2018). An atlas of genetic associations in UK Biobank. Nature Genetics, 50(11), 1593–1599. https://doi.org/10.1038/s41588-018-0248-z

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Frontiers in Genetics, 11, 424. https://doi.org/10.3389/fgene.2020.00424

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dutta, P., Talenti, A., Young, R., Jayaraman, S., Callaby, R., Jadhav, S. K., Dhanikachalam, V., Manikandan, M., Biswa, B. B., Low, W. Y., Williams, J. L., Cook, E., Toye, P., Wall, E., Djikeng, A., Marshall, K., Archibald, A. L., Gokhale, S., Kumar, S., … Prendergast, J. G. D. (2020). Whole genome analysis of water buffalo and global

cattle breeds highlights convergent signatures of domestication. Nature Communications, 11(1), 4739. https://doi.org/10.1038/s41467-020-18550-1

Edinburgh Compute and Data Facility web site. (2021). U of Edinburgh. <www.ecdf.ed.ac.uk>

GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science (New York, N.Y.), 369(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776

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

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., … Kent, W. J. (2006). The UCSC Genome Browser Database: Update 2006. Nucleic Acids Research, 34 D590-598. https://doi.org/10.1093/nar/gkj144

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., & Eskin, E. (2014). Identifying Causal Variants at Loci with Multiple Signals of Association. Genetics, 198(2), 497–508. https://doi.org/10.1534/genetics.114.167908

Käser, T. (2021). Swine as biomedical animal model for T-cell research—Success and potential for transmittable and non-transmittable human diseases. Molecular Immunology, 135, 95–115. https://doi.org/10.1016/j.molimm.2021.04.004

Landrum, M. J., & Kattman, B. L. (2018). ClinVar at five years: Delivering on the promise. Human Mutation, 39(11), 1623–1630. https://doi.org/10.1002/humu.23641

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Research, 46(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Lean, I. J., & Rabiee, A. R. (2011). Effect of feeding biotin on milk production and hoof health in lactating dairy cows: A quantitative assessment. Journal of Dairy Science, 94(3), 1465–1476. https://doi.org/10.3168/jds.2010-3682

Li, C., Tian, D., Tang, B., Liu, X., Teng, X., Zhao, W., Zhang, Z., & Song, S. (2021). Genome Variation Map: A worldwide collection of genome variations across multiple species. Nucleic Acids Research, 49(D1), D1186–D1191. https://doi.org/10.1093/nar/gkaa1005

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009).

The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, S., Gao, Y., Canela-Xandri, O., Wang, S., Yu, Y., Cai, W., Li, B., Xiang, R., Chamberlain, A. J., Pairo-Castineira, E., D'Mellow, K., Rawlik, K., Xia, C., Yao, Y., Navarro, P., Rocha, D., Li, X., Yan, Z., Li, C., … Fang, L. (2021). A comprehensive catalogue of regulatory variants in the cattle transcriptome (p. 2020.12.01.406280). bioRxiv. https://doi.org/10.1101/2020.12.01.406280

Low, W. Y., Tearle, R., Bickhart, D. M., Rosen, B. D., Kingan, S. B., Swale, T., Thibaud-Nissen, F., Murphy, T. D., Young, R., Lefevre, L., Hume, D. A., Collins, A., Ajmone-Marsan, P., Smith, T. P. L., & Williams, J. L. (2019). Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. Nature Communications, 10(1), 260. https://doi.org/10.1038/s41467-018-08260-0

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

Lunney, J. K., Van Goor, A., Walker, K. E., Hailstock, T., Franklin, J., & Dai, C. (2021). Importance of the pig as a human biomedical model. Science Translational Medicine, 13(621), eabd5758. https://doi.org/10.1126/scitranslmed.abd5758

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics, 26(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biology, 17(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Meurens, F., Summerfield, A., Nauwynck, H., Saif, L., & Gerdts, V. (2012). The pig: A model for human infectious diseases. Trends in Microbiology, 20(1), 50–57. https://doi.org/10.1016/j.tim.2011.11.002

Navarro Gonzalez, J., Zweig, A. S., Speir, M. L., Schmelter, D., Rosenbloom, K. R., Raney, B. J., Powell, C. C., Nassar, L. R., Maulding, N. D., Lee, C. M., Lee, B. T., Hinrichs, A. S., Fyfe, A. C., Fernandes, J. D., Diekhans, M., Clawson, H., Casper, J., Benet-Pagès, A., Barber, G. P., … Kent, W. J. (2021). The UCSC Genome Browser database: 2021 update. Nucleic Acids Research, 49(D1), D1046–D1057. https://doi.org/10.1093/nar/gkaa1070

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825--2830.

Pir, M. S., Bilgin, H. I., Sayici, A., Coşkun, F., Torun, F. M., Zhao, P., Kang, Y., Cevik, S., & Kaplan, O. I. (2022). ConVarT: A search engine for matching human genetic variants with variants from non-human species. Nucleic Acids Research, 50(D1), D1172–D1178. https://doi.org/10.1093/nar/gkab939

Plassais, J., Kim, J., Davis, B. W., Karyadi, D. M., Hogan, A. N., Harris, A. C., Decker, B., Parker, H. G., & Ostrander, E. A. (2019). Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nature Communications, 10(1), 1489. https://doi.org/10.1038/s41467-019-09373-w

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research, 20(1), 110–121. https://doi.org/10.1101/gr.097857.109

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

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: A next generation web server for deep-sequencing data analysis. Nucleic Acids Research, 44(W1), W160–W165. https://doi.org/10.1093/nar/gkw257

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics, 19(8), 491–504. https://doi.org/10.1038/s41576-018-0016-z

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research, 15(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C., Oksvold, P., Edfors, F., Limiszewska, A., Hikmet, F., Huang, J., Du, Y., Lin, L., Dong, Z., Yang, L., Liu, X., Jiang, H., Xu, X., Wang, J., … Mulder, J. (2020). An atlas of the protein-coding genes in the human, pig, and mouse brain. Science, 367(6482), eaay5947. https://doi.org/10.1126/science.aay5947

Talenti, A., & Prendergast, J. (2021). nf-LO: A Scalable, Containerized Workflow for Genome-to-Genome Lift Over. Genome Biology and Evolution, 13(9), evab183. https://doi.org/10.1093/gbe/evab183

The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G. R., Steering committee, Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S.

B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. https://doi.org/10.1038/nature15393

Walters, E. M., & Prather, R. S. (2013). Advancing Swine Models for Human Health and Diseases. Missouri Medicine, 110(3), 212–215.

Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A. P., van de Geijn, B., Reshef, Y., Márquez-Luna, C., O'Connor, L., Pirinen, M., Finucane, H. K., & Price, A. L. (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nature Genetics, 52(12), 1355–1363. https://doi.org/10.1038/s41588-020-00735-5

Wen, X., Pique-Regi, R., & Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genetics, 13(3), e1006646. https://doi.org/10.1371/journal.pgen.1006646

Wilkin, D. J., Szabo, J. K., Cameron, R., Henderson, S., Bellus, G. A., Mack, M. L., Kaitila, I., Loughlin, J., Munnich, A., Sykes, B., Bonaventure, J., & Francomano, C. A. (1998). Mutations in fibroblast growth-factor receptor 3 in sporadic cases of achondroplasia occur exclusively on the paternally derived chromosome. American Journal of Human Genetics, 63(3), 711–716. https://doi.org/10.1086/302000

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., … Flicek, P. (2020). Ensembl 2020. Nucleic Acids Research, 48(D1), D682–D688. https://doi.org/10.1093/nar/gkz966

Zhu, F., Nair, R. R., Fisher, E. M. C., & Cunningham, T. J. (2019). Humanising the mouse genome piece by piece. Nature Communications, 10(1), 1845. https://doi.org/10.1038/s41467-019-09716-7

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

Ziegler, A., Gonzalez, L., & Blikslager, A. (2016). Large Animal Models: The Key to Translational Discovery in Digestive Disease Research. Cellular and Molecular Gastroenterology and Hepatology, 2(6), 716–724. https://doi.org/10.1016/j.jcmgh.2016.09.003