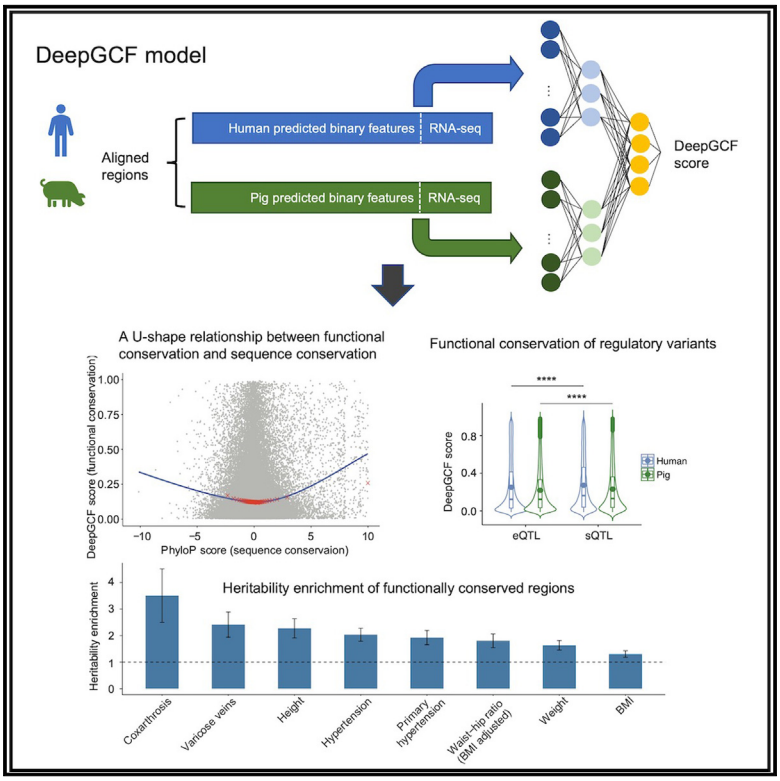
Cell Genomics

Article

Learning functional conservation between human and pig to decipher evolutionary mechanisms underlying gene expression and complex traits

Graphical abstract



Highlights

- DeepGCF improves the prediction accuracy of functional conservation
- Sequence conservation shows a U-shaped relationship with functional conservation
- Functionally conserved regions play key roles in regulatory activities
- Functionally conserved regions show heritability enrichment in human complex traits

Authors

Jinghui Li, Tianjing Zhao, Dailu Guan, ..., Huaijun Zhou, Lingzhao Fang, Hao Cheng

Correspondence

lingzhao.fang@qgg.au.dk (L.F.), qtlcheng@ucdavis.edu (H.C.)

In brief

Li et al. developed a deep learning model, DeepGCF, to learn the genomic conservation at the functional level between human and pig using epigenome and gene expression profiles. They identified a core set of regions as functionally conserved that plays key roles in gene regulation and complex traits in humans.

Li et al., 2023, Cell Genomics *3*, 100390
October 11, 2023
https://doi.org/10.1016/j.xgen.2023.100390

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Learning functional conservation between human and pig to decipher evolutionary mechanisms underlying gene expression and complex traits

Jinghui Li,[1] Tianjing Zhao,[1] Dailu Guan,[1] Zhangyuan Pan,[1] Zhonghao Bai,[2] Jinyan Teng,[3] Zhe Zhang,[3] Zhili Zheng,[4,5] Jian Zeng,[4] Huaijun Zhou,[1] Lingzhao Fang,[2,*] and Hao Cheng[1,6,*]

[1]Department of Animal Science, University of California, Davis, Davis, CA 95616, USA
[2]Center for Quantitative Genetics and Genomics (QGG), Aarhus University, 8000 Aarhus, Denmark
[3]State Key Laboratory of Swine and Poultry Breeding Industry, National Engineering Research Center for Breeding Swine Industry, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China
[4]Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia
[5]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[6]Lead contact
*Correspondence: lingzhao.fang@qgg.au.dk (L.F.), qtlcheng@ucdavis.edu (H.C.)
https://doi.org/10.1016/j.xgen.2023.100390

SUMMARY

Assessment of genomic conservation between humans and pigs at the functional level can improve the potential of pigs as a human biomedical model. To address this, we developed a deep learning-based approach to learn the genomic conservation at the functional level (DeepGCF) between species by integrating 386 and 374 functional profiles from humans and pigs, respectively. DeepGCF demonstrated better prediction performance compared with the previous method. In addition, the resulting DeepGCF score captures the functional conservation between humans and pigs by examining chromatin states, sequence ontologies, and regulatory variants. We identified a core set of genomic regions as functionally conserved that plays key roles in gene regulation and is enriched for the heritability of complex traits and diseases in humans. Our results highlight the importance of cross-species functional comparison in illustrating the genetic and evolutionary basis of complex phenotypes.

INTRODUCTION

Comparative genomics not only reveals evolutionary changes at the DNA sequence level[1] but also helps with translating genetic and biological findings across species. Compared with model laboratory organisms like mice, pigs (*Sus scrofa*) are more similar to humans in terms of anatomy, physiology, and gene-regulatory mechanisms,[2] making them biomedical and genetic models for human medicine and genetic diseases, including studies of drugs, xenotransplantation, Alzheimer's disease, breast cancer, and diabetes.[3–7] To fully recognize the substantial potential of pigs as a human biomedical model, it is essential to conduct extensive comparisons of pig and human physiology at the molecular level and to assess the degree to which genetic and biological findings in pigs can be extrapolated to humans. Methods have been proposed to infer conservation at the DNA sequence level, such as genomic evolutionary rate profiling (GERP) and phylogenetic p values (PhyloP).[8,9] However, conservation at the DNA sequence level does not necessarily reflect conservation at the functional level.[10–12]

The ongoing global efforts on functional annotation of genomes in humans and livestock, such as the Encyclopedia of DNA Elements,[13] Roadmap Epigenomics,[14] Functional Annotation of Animal Genomes (FAANG),[15] and Farm Animal Genotype-Tissue Expression (FarmGTEx) projects,[16] provide an opportunity to quantify functional genomic conservation across species. Previous studies have often relied on a single functional profile in one tissue/cell type, such as gene expression or a specific epigenetic mark, to infer the functional conservation of orthologous regions between humans and pigs.[17–19] However, integrative analyses of multi-omics measurements are needed to unravel how biological information encoded in the genome is conserved or diverged during evolution. This is because the functional consequence of genomic variants is often modulated at multiple levels of gene regulation across tissues/cells. Artificial neural networks have been applied to predict and integrate multi-omics data, such as histone marks, transcription factors, and gene expression, to investigate transcriptional and biochemical impacts of DNA sequences and their conservation across species.[20,21] For instance, the neural network model, Learning Evidence of Conservation from Integrated Functional genomic annotaions (LECIF), was developed to study human-mouse functional conservation based on multi-omics data from the Roadmap and the Encyclopedia of DNA Elements (ENCODE) databases.[21]

Here, we developed a deep learning-based approach called DeepGCF (genomic conservation at the functional level) to systematically evaluate the functional conservation between humans

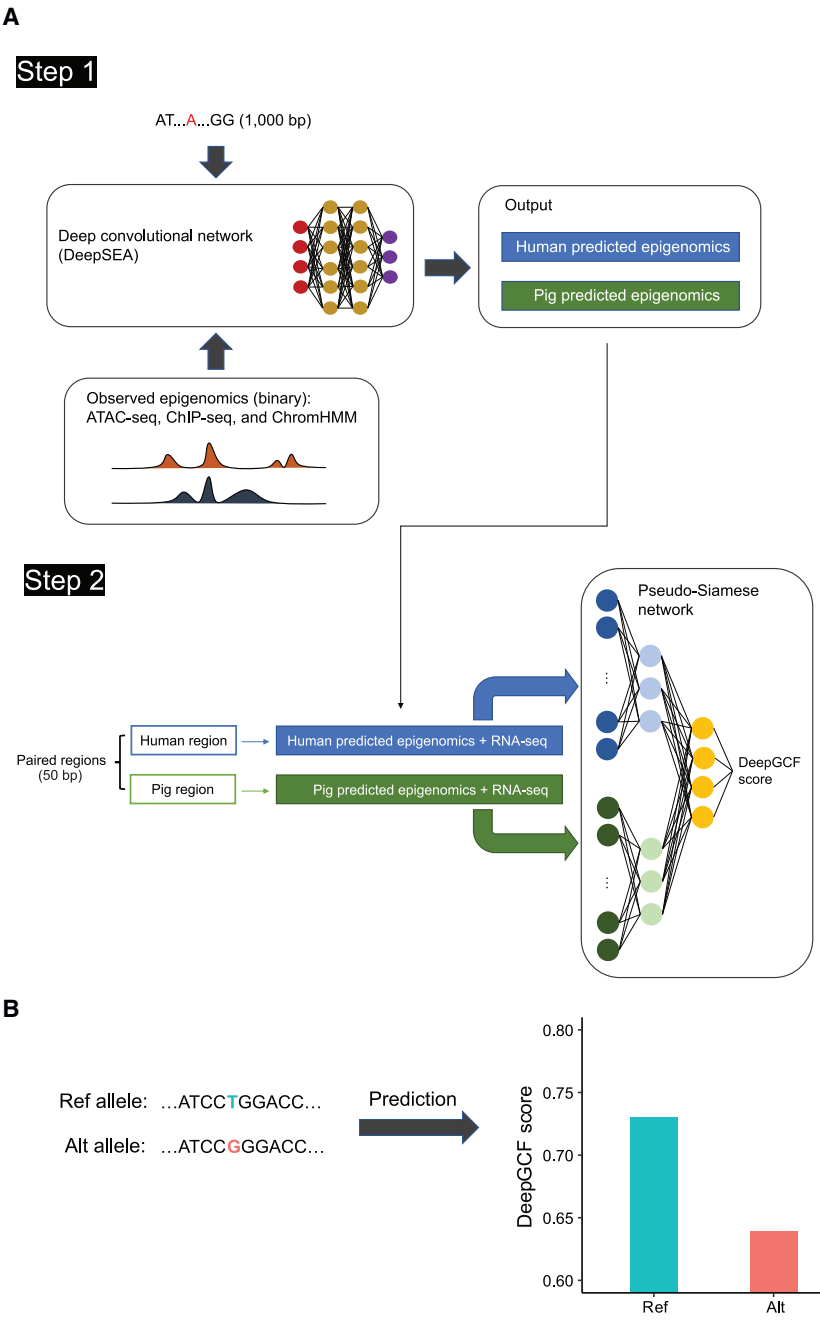


Figure 1. Overview of the DeepGCF model

(A) The learning procedure of the DeepGCF model consists of two steps. The first step is to train DeepSEA models in humans and pigs separately to transform the binary functional features (e.g., peaks called from ATAC-seq and ChIP-seq and chromatin states predicted from a multivariate Hidden Markov Model (ChromHMM)) to continuous values by predicting the functional effects of single nucleotides through centering the target nucleotide at a genomic region of 1,000 bp. The second step is to train a pseudo-Siamese network to predict whether the paired human-pig regions are orthologous using two corresponding vectors of functional effects predicted from DeepSEA and normalized gene expression as input. The output, DeepGCF score, is a value between 0 and 1 quantifying the functional conservation of the paired human-pig region.

(B) The DeepGCF model can be applied to predict the effect of genome variants on functional conservation, quantified by changes in DeepGCF scores.

and pigs. Unlike LECIF, which uses functional genomics data as input, DeepGCF incorporates DNA sequences and functional genomics data as input. This enables us to predict the impact of sequence mutations on the functional conservation between species. By integrating 386 epigenome and transcriptome profiles from 28 tissues in humans and 374 epigenome and transcriptome profiles from 21 tissues in pigs, DeepGCF captures the functional conservation of epigenetic features and genes across tissues between humans and pigs.

Furthermore, we examined expression/splicing quantitative trait loci (e/sQTLs) from 49 human GTEx tissues and 34 PigGTEx tissues[22,23] as well as integrated cross-species comparisons of the results from genome-wide association studies (GWASs) of 80 complex traits/diseases in humans. Deep GCF provides novel insights into the evolutionary mechanisms underlying molecular and complex phenotypes. The DeepGCF model can be expanded to more than two species to understand the evolution of the functional genome as large-scale functional annotation data become available for multiple species in the near future.

RESULTS

Overview of the DeepGCF model

Training of the DeepGCF model consists of two steps (Figure 1). The first step converts binary functional features to continuous values by training a deep convolutional network implemented in the deep learning-based sequence analyzer (DeepSEA).[24] Binary functional features are commonly used in functional genomics to represent whether a genomic base overlaps with functional annotations, such as peaks or chromatin states obtained from an assay for transposase-accessible chromatin sequencing (ATAC-seq) and chromatin immunoprecipitation sequencing (ChIP-seq) experiments. DeepSEA takes DNA sequences and binary functional features as input and predicts the probabilities of each functional feature at single-nucleotide resolution. In this study, we collected 309 and 294 genome-wide binary functional annotations from humans and pigs, respectively (Tables S1–S4).

The functional annotations represented chromatin accessibility measured by ATAC-seq, histone modifications measured by ChIP-seq, and predicted chromatin states from 26 and 21 tissues

in humans and pigs, respectively. The human ATAC-seq and ChIP-seq data were obtained from ENCODE,[13] while those of pigs were from Pan et al.[18] and Zhao et al.[18,19] The predicted chromatin states of humans and pigs were obtained from Pan et al.[18] We trained the DeepSEA models and predicted the functional effect of each nucleotide in humans and pigs separately, which was subsequently used as input for the DeepGCF model to predict the functional conservation score between these two species. The performance of DeepSEA was evaluated with an independent validation set and showed predictive power for both species (Figure S1).

The second step of DeepGCF predicts the functional conservation score of orthologous regions between humans and pigs with a supervised deep learning approach, similar to LECIF.[21] A whole-genome alignment between humans and pigs was divided into non-overlapping 50-bp regions within each alignment block, resulting in 38,961,848 paired alignments (i.e., orthologous regions), covering ~42% of the entire human genome. The first base of each 50-bp region was selected to represent the functional annotation of the entire region because bases in such narrow regions are likely to have similar functions, and this reduces the computational burden.[21] In addition to the predicted functional effects from DeepSEA, we included gene expression values from 77 and 80 RNA sequencing (RNA-seq) datasets as functional annotations, representing 11 and 19 tissues in humans and pigs, respectively (Tables S5 and S6).[13,18,19] To train the DeepGCF model, we randomly shifted the human-pig orthologous regions to obtain an equal number of non-orthologous pairs. Because there is a lack of ground truth for predicted functional conservation in the absence of relevant experimental data, we approximated that orthologous regions (coded as 1) are more likely to be functionally conserved than non-orthologous ones (coded as 0). We then trained a pseudo-Siamese neural network model using functional effects predicted from DeepSEA and gene expression as input (Figure 1A).[25]

During model training, non-orthologous regions were weighted 50 times more than orthologous ones to emphasize regions with strong evidence of functional conservation.[21] The output of the model, the DeepGCF score, is a value between 0 and 1 that quantifies the functional conservation of the paired human-pig region. Furthermore, because DeepGCF predicts the functional conservation from DNA sequences, it allows us to conduct *in silico* mutagenesis analysis. This analysis assesses the impact of orthologous variants on functional conservation between species by investigating changes in the DeepGCF score caused by a genetic mutation (Figure 1B).

Evaluation of the DeepGCF model

The performance of DeepGCF was evaluated with an independent testing set to predict whether paired human-pig regions are orthologous. DeepGCF showed a better predictive ability compared with LECIF, with areas under the receiver operating characteristic curve (AUROC) and precision-recall curve (AUPRC) of 0.89 and 0.87, respectively, while LECIF had an AUROC and AUPRC of 0.80 and 0.79, respectively (Figures 2A and 2B). Among all orthologous regions between humans and pigs, only a small percentage (1.2%) had a DeepGCF score greater than 0.8, while more than half had a score of less than 0.1 (Figure 2C). These results indicate that most orthologous regions were not functionally conserved between these two species, consistent with previous findings for humans and mice.[21]

Notably, to make the number of functional features comparable between pigs and humans, we only collected the human functional profiles at the tissue level. Furthermore, we merged multiple binary functional profiles of the same type from the same tissue into one profile to reduce the computational load. This resulted in 386 and 374 functional features in humans and pigs, respectively. In addition, we tested the performance of DeepGCF using all 861 human profiles and 577 pig profiles without merging the binary functional profiles. The result showed that a DeepGCF model that used all functional profiles had a consistent prediction accuracy compared with a model trained with merged datasets (Figure S2). We also normalized the gene expression values with a natural logarithm transformation, which resulted in a better prediction accuracy compared with one without transformation (Figure S2).

We further explored features that may influence the model's performance, including sample size and diversity of functional annotations regarding array and tissue/cell type. Downsampling of functional profiles in humans and pigs during model training indicates that analyses that use approximately 50% (human, 192; pig, 187) or 10% (human, 52; pig, 47) of the currently available profiles resulted in similar AUROC (50%, 0.88; 10%, 0.85) and AUPRC (50%, 0.87; 10%, 0.83) values compared with tests that use all available profiles. However, using only approximately 1% (human, 4; pig, 4) of the profiles resulted in substantially lower AUROC (0.69) and AUPRC (0.68) values (Figure 2D). When leaving one type of functional profiles out, the predictive ability of DeepGCF remained similar (Figure 2E).

Relationship between DNA sequence conservation and functional conservation

To explore whether DNA sequence conservation indicates functional conservation, we investigated the relationship between the DeepGCF score and the PhyloP score, a commonly used measure of the DNA sequence conservation across species.[9] We observed a U-shaped relationship between the PhyloP scores and the DeepGCF scores (Figure 3A). This suggests that rapidly and slowly evolving sequences exhibited a higher functional conservation between species compared with sequences that are evolutionarily neutral or nearly neutral. This finding is consistent with comparisons of individual epigenetic marks and DNA sequence conservation.[18,26]

We defined three groups of orthologous regions from their PhyloP and DeepGCF scores, representing the two tails and the bottom of the U curve: (1) regions with high DeepGCF (>95th percentile) and PhyloP (>95th percentile), referred to as high D & high P (n = 260,281); (2) regions with high DeepGCF (>95th percentile) but low PhyloP (<5th percentile), referred to as high D & low P (n = 152,557); and (3) regions with low DeepGCF (<5th) and medium PhyloP (between 47.5th and 52.5th), referred to as low D & med P (n = 95,231).

We then examined sequence classes and Gene Ontology (GO) terms for these three groups of regions. We determined sequence classes from predicted regulatory activities of DNA sequences in the human genome using a deep learning model, Sei, trained on a compendium of 21,907 epigenome profiles.[27] We found that high D & high P regions were enriched in sequences

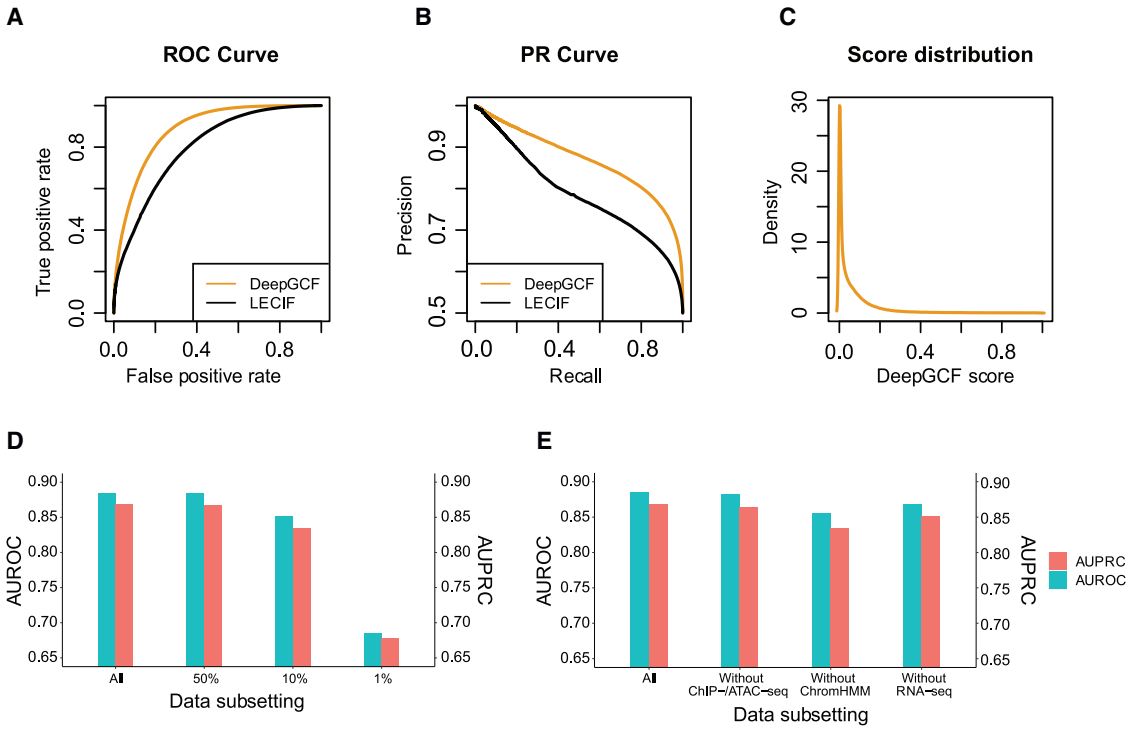


Figure 2. The performance of DeepGCF under different scenarios

(A) Receiver operating characteristic (ROC) curves comparing the performance of DeepGCF (this study) and LECIF[21] methods. The ROC curve of each method is generated by predicting whether 200,000 pairs randomly selected from the testing set, which included equal numbers of orthologous and non-orthologous pairs, were orthologous.

(B) Precision-recall (PR) curves generated by similar procedures as the ROC curves.

(C) The distribution of DeepGCF scores across all 38,961,848 human-pig ortholog pairs.

(D) The areas under the ROC curve (AUROC) and PR curve (AUPRC) of DeepGCF using all (human, 386; pig, 374), ~50% (human, 192; pig, 187), ~10% (human, 52; pig: 47), and ~1% (human, 4; pig: 4) of human and pig functional features. The subsets of the human and pig features were randomly and proportionally selected from each of the ChIP-seq/ATAC-seq, ChromHMM, and RNA-seq profiles.

(E) The AUROC and AUPRC of DeepGCF using all functional features (human, 386; pig, 374), features without ChIP-seq/ATAC-seq (human, 129; pig, 84), features without ChromHMM (human, 180; pig, 210), and features without RNA-seq (human, 77; pig, 80).

with a predicted promoter, CTCF binding sites, and transcriptional effects but depleted in enhancer regions relative to the whole genome (binomial test, $p < 2.2 \times 10^{-16}$; Figure 3B). High D & high P regions showed more enrichment in transcription compared with other regions (binomial test, $p < 2.2 \times 10^{-16}$; Figure 3B) and were significantly associated with RNA-related regulation processes (binomial test, false discovery rate [FDR] < 0.05; hypergeometric test, FDR <0.05; Table S7), indicating similarities in transcriptional networks between pigs and humans.[17,28] High D & low P regions were significantly enriched in Polycomb (binomial test, $p < 2.2 \times 10^{-16}$; Figure 3B), which is consistent with the fact that some core subunits of Polycomb protein complexes with similar biological functions have shown weak evolutionary conservation in DNA sequence across species.[29] Low D & medium P regions had similar sequence class compositions as the whole genome, with the exception of promoter regions, which were enriched, but to a lesser extent than high D & high P and high D & low P (binomial test, $p < 2.2 \times 10^{-16}$; Figure 3B). Low D & med P regions were also enriched in fewer GO terms than regions with high DeepGCF scores (Tables S7–S9). In addition, we examined six different sequence ontologies and found that the 5′ UTR is the most functionally conserved element, followed by the start codon, 3′ UTR, stop codon, exon, and finally intron. This finding is consistent between humans and pigs (Figure 3C).

To investigate the impact of orthologous variants on functional conservation, we examined 35,575,835 human SNPs that are located in orthologous regions between humans and pigs as ascertained in the 1000 Genomes Project.[30] We used the DeepGCF model, which was trained exclusively on the predicted probabilities of binary features from DeepSEA (i.e., leaving RNA-seq out) because the DeepSEA model does not predict continuous functional features. The new score predicted from DeepGCF without RNA-seq data showed a relatively good agreement with the original DeepGCF score, with a Pearson's correlation coefficient (PCC) of 0.74 (Figure S3).

To measure the effect of each human SNP on functional conservation, we recomputed the probabilities of binary features for the corresponding orthologous human region because of the SNP mutation while keeping the pig probabilities the same and then used the new probabilities to calculate the updated DeepGCF score. The effect on functional conservation is

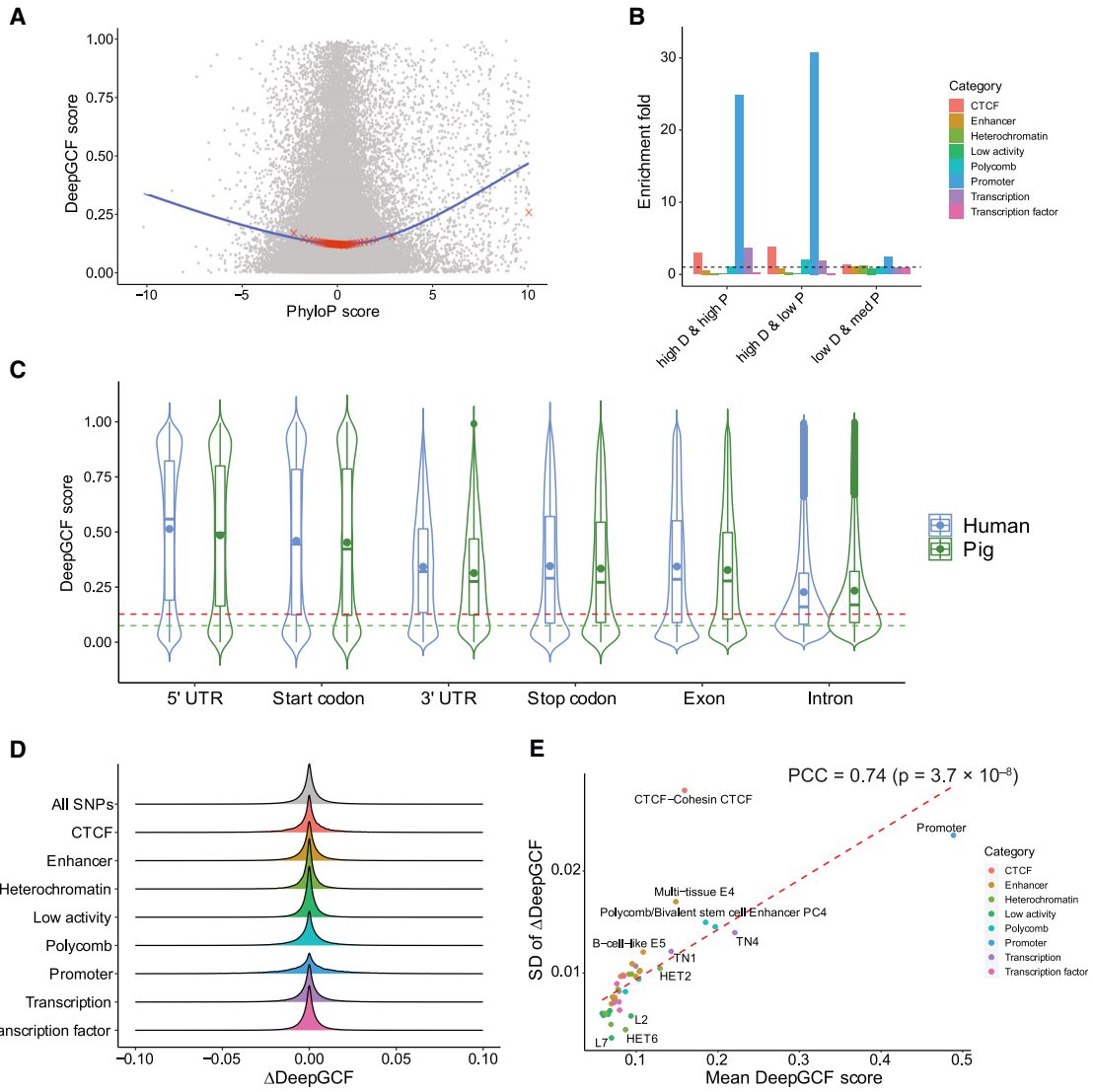


Figure 3. Comparison of functional and sequence conservations

(A) Relationship between DeepGCF scores and PhyloP scores of 20,000 randomly selected human regions. The PhyloP score is based on multiple alignments of 99 vertebrate genomes to the human genome.[9] The blue line is the fitted loess regression. The red crosses represent 50 equally divided percentiles of the PhyloP score and corresponding mean DeepGCF score.

(B) Enrichment fold of 8 sequence class categories[27] for regions with high DeepGCF (>95th percentile) and high PhyloP (>95th percentile, high D & high P, n = 260,281) and regions with high DeepGCF (<5th percentile) and medium PhyloP (between 47.5th and 52.5th percentile, low D & med P, n = 77,848). Enrichment is equal to the proportion of a sequence class category for a type of orthologous region divided by that for the whole genome. The dashed line (= 1) represents no enrichment.

(C) Distribution of DeepGCF score for different sequence ontologies. The red and green dashed lines represent the mean and median DeepGCF score of the whole genome, respectively. The dots in each box represent the mean DeepGCF score. In each box, the center line represents the median, the dot represents the mean, box limits represent the upper and lower quartiles, whiskers represent 1.5 × interquartile range, and individual points are outliers.

(D) ΔDeepGCF (DeepGCF after mutation – original DeepGCF) caused by 1,000,000 randomly selected orthologous variants, which are classified into 8 sequence class categories annotated by Sei.[27].

(E) The effect of orthologous variants (n = 35,575,835) on the DeepGCF score of regions in 40 sequence classes annotated by Sei,[27] which are classified into 8 categories. The effect was measured by ΔDeepGCF for variants in each sequence class. The SD of ΔDeepGCF for each sequence class quantifies the sensitivity of the sequence class to variants. The dashed line is the fitted regression line.

measured by ΔDeepGCF = DeepGCF after SNP mutation – original DeepGCF. By classifying all orthologous variants into eight sequence class categories,[27] we found that most variants had a limited effect on functional conservation (Figure 3D). We further grouped them into 40 sequence classes[27] and found that genetic mutations in sequence classes with higher DeepGCF scores (more functionally conserved) are more likely to have larger impacts (SD of ΔDeepGCF) on functional conservation between

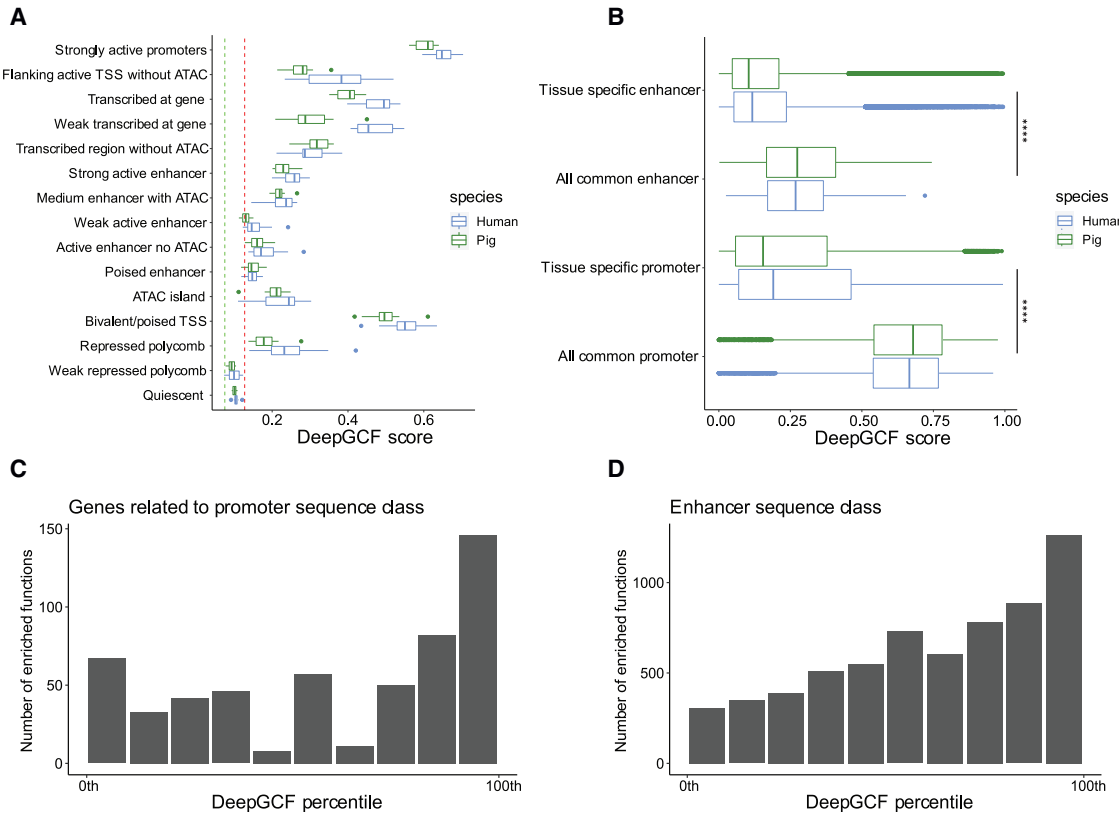


Figure 4. DeepGCF scores of genomic regions overlapping with regulatory elements

(A) Distribution of average DeepGCF scores across human tissues (n = 12) and pig tissues (n = 14) for each chromatin state. The red and green dashed lines represent the mean and median DeepGCF score of the whole genome. In each box, the center line represents the median, box limits represent the upper and lower quartiles, whiskers represent 1.5 × interquartile range, and individual points are outliers.

(B) DeepGCF scores of genomic regions overlapping with tissue-specific strongly active promoters and enhancers for human and pig.[18] "All common" represents promoters/enhancers shared across all tissues. Asterisks denote two-sided Mann-Whitney U test: ****$p < 2.2 \times 10^{-16}$.

(C) Number of significantly enriched GO terms for human of genes related to promoters annotated by Sei.[27] Significance was calculated using FDR < 0.05 for the binomial and hypergeometric tests. The genes were binned by DeepGCF into 10 equal-width bins, and a functional enrichment analysis was conducted on each bin.

(D) Similar to (C) but showing the results of enhancers annotated by Sei.[27].

species (Figure 3E). Notably, the average DeepGCF score of CTCF binding sites is lower than that of promoters but more sensitive to genetic mutations, indicating that genetic disruption of CTCF binding sites had a pronounced impact on functional genome evolution between species by altering genome topology and gene expression.[31,32]

DeepGCF captures the evolutionary characteristics of regulatory elements

To investigate the functional conservation of distinct regulatory elements between pigs and humans, we examined the DeepGCF score of 15 chromatin states predicted in 14 pig tissues and 12 human tissues.[18] We found that strongly active promoters had the highest DeepGCF scores, indicating the strongest functional conservation, followed by a poised transcription start site (TSS), chromatin states proximal to the TSS, enhancers, and, finally, repressed Polycomb (Figure 4A).This conservation pattern was consistent between humans and pigs, which aligns with previous studies on the conservation properties of regulatory elements.[18,33]

Because tissues may have specific chromatin states that play crucial roles in determining cellular functions, we identified strongly active promoters and enhancers that were tissue specific in each of 12 human tissues and 14 pig tissues. Compared with promoters and enhancers shared by all tissues, tissue-specific ones showed significantly lower DeepGCF scores in both species (Mann-Whitney U test, $p < 2.2 \times 10^{-16}$), suggesting a faster evolutionary rate for tissue-specific regulatory elements (Figure 4B). Among the eight tissues we examined in humans and pigs, we found that adipose tissue exhibited the strongest conservation of promoters in human and pig, followed by spleen, lung, cortex, liver, and stomach tissue (Figure S4A). However, the conservation patterns of enhancers were not consistent between species and varied among tissues (Figure S4B).

We further investigated the DeepGCF score on human promoters and enhancers annotated by Sei.[27] We linked a promoter to its potential target gene and then ranked genes based on the DeepGCF scores of their promoters, from highest to lowest. We observed that the top 5% ranked genes were significantly

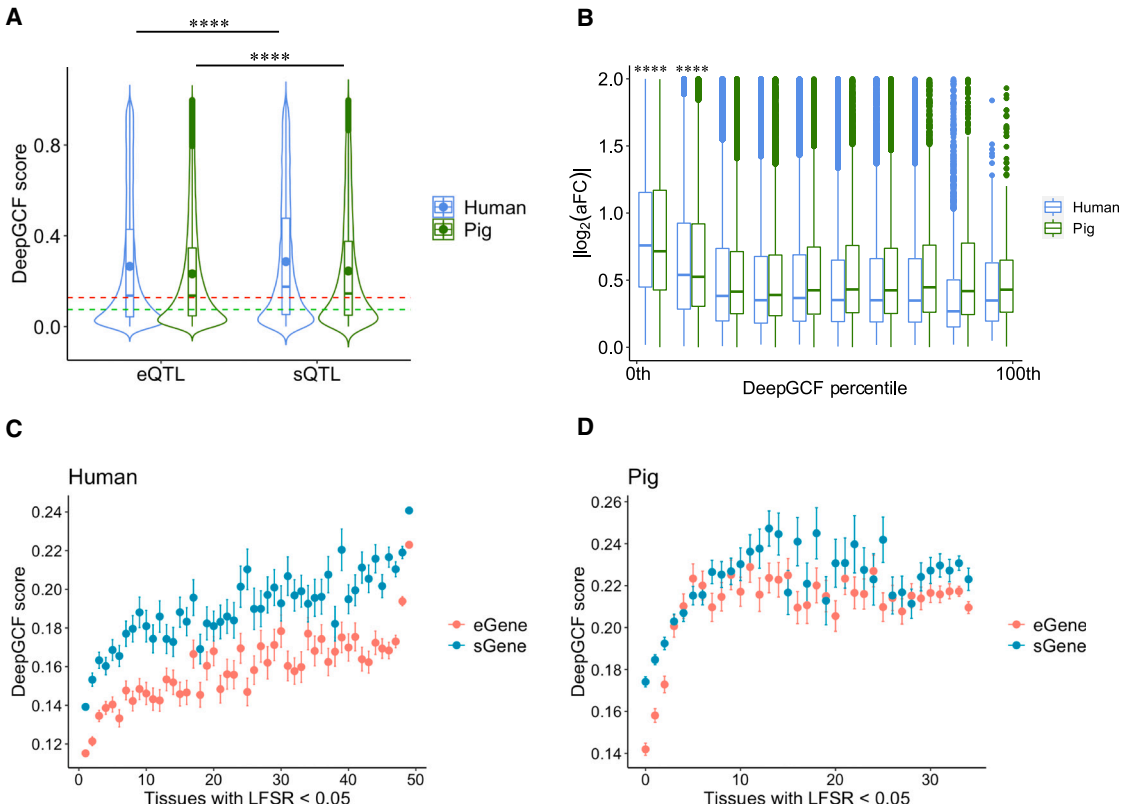


Figure 5. Relationship of DeepGCF scores to genetic variants

(A) The distribution of DeepGCF scores for eQTLs and sQTLs. The red and green dashed lines represent the mean and median DeepGCF score of the whole genome, respectively. Asterisks denote two-sided Mann-Whitney U test: ****$p < 10^{-8}$. In each box, the center line represents the median, the dot represents the mean, box limits represent the upper and lower quartiles, whiskers represent 1.5 × interquartile range, and individual points are outliers.

(B) Relationship between the absolute value of eQTL effect size measured by log allelic fold change (|log2(aFC)|) and DeepGCF score for eGenes. The genes were binned by DeepGCF into 10 equal-width bins for human and pig, respectively. Asterisks denote that the group is different from all other groups: ****$p < 10^{-8}$ based on Tukey's multiple comparisons.

(C) DeepGCF scores of tissue-sharing e/sGenes from human at local false sign rate (LFSR) < 5% obtained by MashR.[36] Each solid line represents ± standard deviation.

(D) Similar to (C) but showing the results for pigs.

enriched in basic biological processes, such as anatomical structure development and organ morphogenesis (binomial test, FDR < 0.05; hypergeometric test, FDR <0.05), whereas the bottom 5% of genes were significantly enriched in biosynthetic and metabolic processes (binomial test, FDR < 0.05; hypergeometric test, FDR < 0.05; Tables S10 and S11). Additionally, we ranked enhancers with DeepGCF scores and investigated the function of the top 5% and bottom 5% enhancers. Unlike promoters, the top 5% of enhancers exhibited the most significant enrichment in metabolic processes (binomial test, FDR < 0.05; hypergeometric test, FDR < 0.05), while the bottom 5% of enhancers were significantly enriched in organ growth and development (binomial test, FDR < 0.05; hypergeometric test, FDR < 0.05; Tables S12 and S13). Overall, we found that promoters and enhancers with higher DeepGCF scores were enriched in a greater number of biological processes compared with those with lower DeepGCF scores (Figures 4C and 4D), which indicates that functionally conserved regions tend to be hotspots of regulatory activities.

DeepGCF provides insight into the functional conservation of regulatory variants

To explore the functional conservation of regulatory variants, we systematically examined eQTLs and sQTLs in orthologous regions of 49 human tissues and 34 pig tissues, respectively.[22,23] DeepGCF scores of eQTLs and sQTLs were significantly higher (Mann-Whitney U test, $p < 2.2 \times 10^{-16}$) than the genome background across all tissues in humans and pigs (Figures 5A, S5, and S6), which suggests that regulatory variants are functionally conserved between species.[34,35] Notably, sQTLs exhibited higher DeepGCF scores than eQTLs in both species (Mann-Whitney U test, $p < 10^{-8}$), consistent with studies that showed that sQTLs were more enriched in the 5′ UTR than eQTLs[22] and that the 5′ UTR is the most functionally conserved genomic feature (Figure 3C).

Genes with eQTLs or sQTLs were called eGenes and sGenes, respectively. We observed that eGenes that have a larger absolute effect on gene expression had lower DeepGCF scores in both species (Tukey's multiple comparisons, $p < 10^{-8}$; Figure 5B).

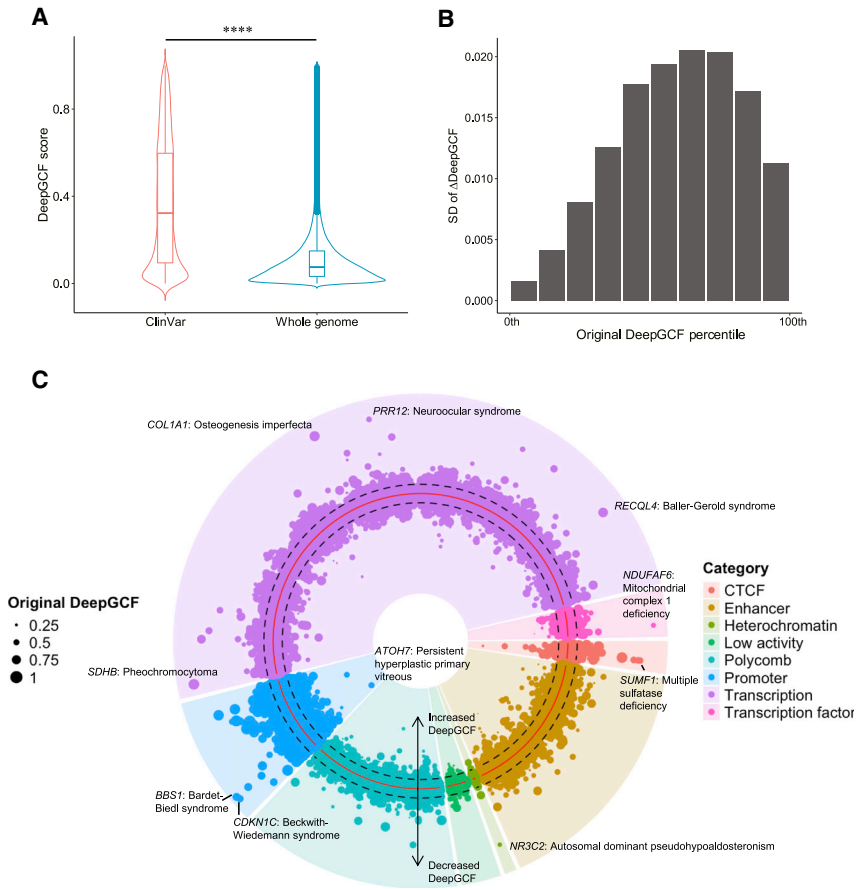


Figure 6. Relationship of conservation score to pathogenic variants

(A) The distribution of DeepGCF scores for pathogenic and likely pathogenic SNPs (n = 104,033) obtained from ClinVar,[38] compared with the distribution of DeepGCF scores across the whole genome. Asterisks denote two-sided Mann-Whitney U test: ****$p < 5 \times 10^{-8}$. In each box, the center line represents the median, box limits represent the upper and lower quartiles, whiskers represent 1.5 × interquartile range, and individual points are outliers.

(B) SD of ΔDeepGCF (DeepGCF after mutation – original DeepGCF) caused by ClinVar SNPs. The SNPs were binned by their original DeepGCF into 10 equal-width bins.

(C) ClinVar SNPs classified by Sei.[27] A polar coordinate system was used, where the radial coordinate indicates the SNP effect on DeepGCF score. The red solid circle represents zero DeepGCF change, and two dashed circles represent ±0.03 of DeepGCF encompassing 95% of SNPs. Each dot represents a SNP, and SNPs in the red circle were predicted to have positive effects (increased DeepGCF), while SNPs outside of the red circle were predicted to have negative effects (decreased DeepGCF). Dot size indicates the original DeepGCF. Within each sequence class, SNPs were ordered by chromosomal coordinates. Diseases and gene names associated with the top 10 SNPs with the largest impact on DeepGCF were annotated.

This observation suggests that orthologous regions with smaller regulatory effects are more likely to be functionally conserved between species, possibly because of stronger purifying selection.[37] Furthermore, regulatory variants influencing more tissues had higher DeepGCF scores, consistent in humans and pigs (Figures 5C and 5D). In addition, the tissue-sharing pattern of orthologous eGenes (PCC = 0.38, $p < 2.2 \times 10^{-16}$) and sGenes (PCC = 0.45, $p < 2.2 \times 10^{-16}$) were positively correlated between humans and pigs. Taken together, these results suggest that regulatory variants controlling transcriptome function in multiple tissues tend to be more functionally conserved between species.

We then investigated the DeepGCF scores of 105,461 pathological and likely pathological SNPs obtained from the ClinVar database.[38] 98.6% of these SNPs were located in the human-pig orthologous regions, consistent with findings that reported more than 98% of pathological variants of Mendelian diseases located in human-mouse orthologous regions.[39] Compared with random orthologous regions, these pathological SNPs were significantly more functionally conserved (Mann-Whitney U test, $p < 2.2 \times 10^{-16}$; Figure 6A).

Similar to orthologous SNPs, we classified the ClinVar SNPs into eight sequence class categories[27] and conducted an *in silico* mutagenesis analysis to predict their impact on functional conservation. The average magnitude of variant effect (measured by |ΔDeepGCF|) for pathological and likely pathological mutations is 1.5 times larger than that for random orthologous SNPs (0.0088 versus 0.0058; Mann-Whitney U test, $p < 2.2 \times 10^{-16}$). The DeepGCF score did not change significantly with specific genetic mutations in most cases, but the variance of ΔDeepGCF showed a bell-shaped curve with respect to the original DeepGCF score. SNPs with medium-high DeepGCF scores (50th–80th percentile) were more sensitive to pathological mutations than those with lower or higher DeepGCF scores (Figure 6B). This suggests that the most functionally conserved regions (>90th percentile) tolerate more mutations than less conserved ones (50th–80th percentile).

The majority of the ClinVar SNPs were classified as transcription (51.2%), followed by enhancer (16.4%), Polycomb (14.8%), promoter (8.8%), transcription factor (3.3%), and CTCF (2.2%; Figure 6C). Among the ClinVar SNPs with top 5% of |ΔDeepGCF| (>0.03), there were more SNPs relevant to a decreased DeepGCF (54.4%) than an increased one (45.6%). Moreover, 9 of 10 ClinVar SNPs with the largest effect on DeepGCF were relevant to a decreased DeepGCF (Figure 6C). In summary, pathological and likely pathological SNPs tend to be located in functionally more conserved regions, and their impact on functional conservation is often related to decreased functional conservation between humans and pigs.

Application of DeepGCF on explaining human complex traits

To investigate whether DeepGCF scores could advance our understanding of the evolutionary basis of complex traits/diseases in humans, we conducted a heritability partitioning analysis using

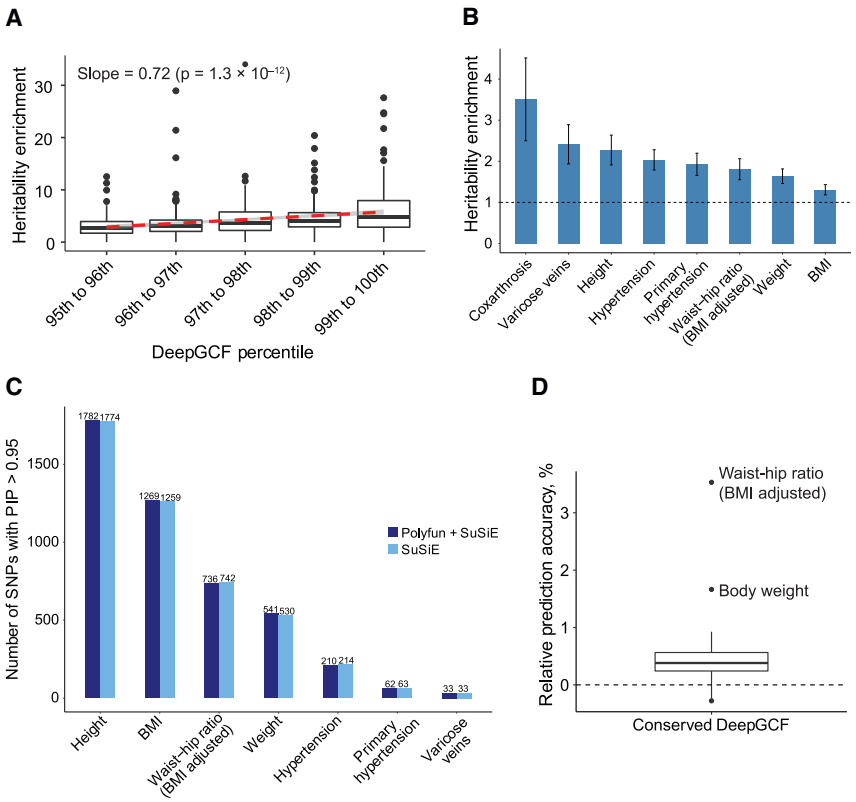


Figure 7. Application of DeepGCF on complex traits/diseases in human

(A) Heritability enrichment calculated by LDSC for 80 human traits using functionally conserved regions (top 5% DeepGCF). The regions were divided into 5 equal equal-width bins, and the heritability enrichment of all traits was calculated for each bin. The red dashed line is the fitted regression line between heritability enrichment and DeepGCF percentile, and the gray area is the 95% confidence interval. In each box, the center line represents the median, box limits represent the upper and lower quartiles, whiskers represent 1.5 × interquartile range, and individual points are outliers.

(B) Significant heritability enrichment (FDR < 0.05) explained by functionally conserved regions for 8 human traits. The error bar is the estimated standard error of heritability enrichment.

(C) The number of putative causal SNPs (PIP > 0.95 and GWAS $p < 5 \times 10^{-8}$) identified by PolyFun + SuSiE[42] with functionally conserved regions as a prior and SuSiE[44] without priors for 7 human traits (the results for coxarthrosis are not shown because no causal SNPs were found using either method).

(D) The relative prediction accuracy of polygenic scores for 20 human complex traits using functionally conserved regions as a prior in SBayesRC.[43] Relative prediction accuracy is equal to (prediction accuracy using the prior − prediction accuracy without priors) / prediction accuracy without priors. Relative prediction accuracy > 0 (dashed line) indicates a higher accuracy than without priors.

the functionally conserved genomic regions (top 5% DeepGCF scores) as a functional annotation to analyze the GWAS summary statistics of 80 human complex traits/diseases (Table S14). This analysis, along with 97 existing annotations from the baseline model of linkage disequilibrium score regression (LDSC),[40,41] indicated that regions with higher DeepGCF scores explained more heritability of complex traits/diseases than those with lower DeepGCF scores (Figure 7A). Specifically, eight complex traits showed a significant heritability enrichment in functionally conserved regions, with the greatest enrichment observed for coxarthrosis (enrichment = 3.5, FDR = 0.032), followed by varicose veins, height, hypertension, primary hypertension, waist-hip ratio, weight, and BMI (Figure 7B; Table S15). Furthermore, we used these eight traits as examples to explore whether DeepCGF could aid fine-mapping of causal variants. We used functionally conserved regions (top 5% of DeepCGF) as a biological prior in the PolyFun + the sum of single effect (SuSiE) model[42] to detect putative causal variants. We found that, compared with the SuSiE model only without any priors, incorporating the functional conservation as a prior led to detection of 33, 22, and 17 additional putative causal variants (posterior inclusion probability (PIP) > 0.95 and $p < 5 \times 10^{-8}$) in height, BMI, and weight, respectively (Figure 7C; Table S16). Additionally, we incorporated functional conservation as a prior in the SBayesRC model[43] to conduct polygenic score prediction for 20 human complex traits (Table S17). On average, the relative prediction accuracy increased by 0.56% (Figure 7D; Table S18), and the largest increase was observed for waist-hip ratio (3.5%), followed by body weight (1.7%). Altogether, our results showed that DeepGCF provides additional insights into the genetic and evolutionary basis of complex phenotypes.

DISCUSSION

In this study, we developed a two-step neural network approach, DeepGCF, to evaluate the genomic conservation at the functional level between humans and pigs. DeepGCF shares a similar model structure as LECIF[21] in evaluation of functional conservation by comparing the epigenome and gene expression profiles of orthologous regions between two species. But instead of using binary epigenome profiles as the direct inputs, DeepGCF first predicts their functional effects (i.e., the continuous probability score of each epigenome binary feature) using DeepSEA[24] and then uses these effects as input to predict the functional conservation between species. Compared with the LECIF approach, DeepSEA showed better performance in ortholog prediction, possibly because of a higher resolution of the model input. Similar to LECIF, we found that the performance of DeepGCF was not sensitive to the number of functional features, indicating that DeepGCF could be applied to other species with fewer functional profiles available. We demonstrated that functional conservation is different from DNA sequence conservation. The relationship between DeepGCF and PhyloP scores confirms a U-shaped relationship between functional and DNA sequence conservation. By examining DeepGCF on chromatin states, sequence ontologies, and regulatory variants, we verified that

DeepGCF captures the functional conservation of the genome and that regions with higher DeepGCF are likely to have more important roles in regulatory activities.

In summary, the DeepGCF approach shows promise as an application for cross-species comparison of functional genomes. We anticipate that the model framework described here can be easily adapted to other species, including humans, mice, pigs, cattle, and other livestock. Generating functional conservation information among different species should provide additional insight into the genetic and evolutionary mechanisms behind complex traits and diseases, analogous to the DNA sequence conservation among vertebrates.

Limitations of the study

Although we expected the DeepGCF to explain genetics of complex traits, the heritability enrichment and polygenic prediction accuracy attributed to functionally conserved regions were limited. This may be because we only considered functional conservation between two species (i.e., humans and pigs) as opposed to multiple species.[45] Because epigenome and gene expression data are generated in other species, we predict an ability to identify the core functionally conserved regions among different evolutionary lineages by expanding the DeepGCF model structure to integrate functional profiles from multiple species. Another limitation is that the functional conservation of the same sequence segment should be conceptually different across different tissues and cell types, which cannot be distinguished by the current DeepGCF score. One ideal way to obtain tissue- and cell-type-specific DeepGCF scores is to train a separate model for each tissue and cell type using the respective data. However, the current volume of functional profiles, particularly in pigs but also for many other vertebrate species, does not support development of tissue- or cell-type-specific DeepGCF models.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Genome alignment
  - Model inputs
  - Prediction of binary functional features based on DeepSEA
  - Data subsets for training and evaluation
  - DeepGCF training
  - Human-pig orthologous SNPs
  - Function enrichment
  - Tissue specific chromatin state
  - Tissue-sharing of e/sGene
  - DeepGCF score for genes
  - Heritability partitioning analysis
  - Fine-mapping analysis
  - Polygenic score prediction
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100390.

ACKNOWLEDGMENTS

We thank Charley Xia and Xiaoshan Yu for helping with the analysis of heritability enrichment and fine-mapping. This work was supported by the USDA Agriculture and Food Research Initiative, National Institute of Food and Agriculture Competitive Grant 2021-67015-33412.

AUTHOR CONTRIBUTIONS

Conceptualization, H.C., L.F., and J.L.; formal analysis, J.L., T.Z., D.G., J.T., and Z. Zheng; data curation, Z.P., Z.B., Z. Zhang, and H.Z.; writing – original draft, J.L.; writing – review & editing, L.F., D.G., J.Z., and H.C.; supervision, H.C. and L.F.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 22, 2023
Revised: April 25, 2023
Accepted: August 2, 2023
Published: August 24, 2023

REFERENCES

1. Alföldi, J., and Lindblad-Toh, K. (2013). Comparative genomics as a tool to understand evolution and disease. Genome Res. *23*, 1063–1068. https://doi.org/10.1101/gr.157503.113.
2. Lunney, J.K., Van Goor, A., Walker, K.E., Hailstock, T., Franklin, J., and Dai, C. (2021). Importance of the pig as a human biomedical model. Sci. Transl. Med. *13*, eabd5758. https://doi.org/10.1126/scitranslmed.abd5758.
3. Schelstraete, W., Devreese, M., and Croubels, S. (2020). Comparative toxicokinetics of Fusarium mycotoxins in pigs and humans. Food Chem. Toxicol. *137*, 111140. https://doi.org/10.1016/j.fct.2020.111140.
4. Montgomery, R.A., Stern, J.M., Lonze, B.E., Tatapudi, V.S., Mangiola, M., Wu, M., Weldon, E., Lawson, N., Deterville, C., Dieter, R.A., et al. (2022). Results of Two Cases of Pig-to-Human Kidney Xenotransplantation. N. Engl. J. Med. *386*, 1889–1898. https://doi.org/10.1056/NEJMoa2120238.
5. Kragh, P.M., Nielsen, A.L., Li, J., Du, Y., Lin, L., Schmidt, M., Bøgh, I.B., Holm, I.E., Jakobsen, J.E., Johansen, M.G., et al. (2009). Hemizygous minipigs produced by random gene insertion and handmade cloning express the Alzheimer's disease-causing dominant mutation APPsw. Transgenic Res. *18*, 545–558. https://doi.org/10.1007/s11248-009-9245-4.
6. Luo, Y., Li, J., Liu, Y., Lin, L., Du, Y., Li, S., Yang, H., Vajta, G., Callesen, H., Bolund, L., and Sørensen, C.B. (2011). High efficiency of BRCA1 knockout using rAAV-mediated gene targeting: developing a pig model for breast cancer. Transgenic Res. *20*, 975–988. https://doi.org/10.1007/s11248-010-9472-8.
7. Renner, S., Braun-Reichhart, C., Blutke, A., Herbach, N., Emrich, D., Streckel, E., Wünsch, A., Kessler, B., Kurome, M., Bähr, A., et al. (2013). Permanent Neonatal Diabetes in INSC94Y Transgenic Pigs. Diabetes *62*, 1505–1511. https://doi.org/10.2337/db12-1065.
8. Cooper, G.M., Stone, E.A., Asimenos, G., NISC Comparative Sequencing Program; Green, E.D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913. https://doi.org/10.1101/gr.3577405.

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

10. Bordeira-Carriço, R., Teixeira, J., Duque, M., Galhardo, M., Ribeiro, D., Acemel, R.D., Firbas, P.N., Tena, J.J., Eufrásio, A., Marques, J., et al. (2022). Multidimensional chromatin profiling of zebrafish pancreas to uncover and investigate disease-relevant enhancers. Nat. Commun. *13*, 1945. https://doi.org/10.1038/s41467-022-29551-7.

11. Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. *42*, 631–634. https://doi.org/10.1038/ng.600.

12. Pennacchio, L.A., and Visel, A. (2010). Limits of sequence and functional conservation. Nat. Genet. *42*, 557–558. https://doi.org/10.1038/ng0710-557.

13. ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640. https://doi.org/10.1126/science.1105136.

14. Roadmap Epigenomics Consortium; Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. https://doi.org/10.1038/nature14248.

15. Andersson, L., Archibald, A.L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W., Casas, E., Cheng, H.H., Clarke, L., Couldrey, C., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. *16*, 57–66. https://doi.org/10.1186/s13059-015-0622-4.

16. Liu, S., Gao, Y., Canela-Xandri, O., Wang, S., Yu, Y., Cai, W., Li, B., Xiang, R., Chamberlain, A.J., Pairo-Castineira, E., et al. (2022). A multi-tissue atlas of regulatory variants in cattle. Nat. Genet. *54*, 1438–1447. https://doi.org/10.1038/s41588-022-01153-5.

17. Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C., Oksvold, P., Edfors, F., Limiszewska, A., Hikmet, F., et al. (2020). An atlas of the protein-coding genes in the human, pig, and mouse brain. Science *367*, eaay5947. https://doi.org/10.1126/science.aay5947.

18. Pan, Z., Yao, Y., Yin, H., Cai, Z., Wang, Y., Bai, L., Kern, C., Halstead, M., Chanthavixay, G., Trakooljul, N., et al. (2021). Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. Nat. Commun. *12*, 5848. https://doi.org/10.1038/s41467-021-26153-7.

19. Zhao, Y., Hou, Y., Xu, Y., Luan, Y., Zhou, H., Qi, X., Hu, M., Wang, D., Wang, Z., Fu, Y., et al. (2021). A compendium and comparative epigenomics analysis of cis-regulatory elements in the pig genome. Nat. Commun. *12*, 2217. https://doi.org/10.1038/s41467-021-22448-x.

20. Wong, A.K., Sealfon, R.S.G., Theesfeld, C.L., and Troyanskaya, O.G. (2021). Decoding disease: from genomes to networks to phenotypes. Nat. Rev. Genet. *22*, 774–790. https://doi.org/10.1038/s41576-021-00389-x.

21. Kwon, S.B., and Ernst, J. (2021). Learning a genome-wide score of human–mouse conservation at the functional genomics level. Nat. Commun. *12*, 2495. https://doi.org/10.1038/s41467-021-22653-8.

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

23. Consortium, T.F.-P., Teng, J., Gao, Y., Yin, H., Bai, Z., Liu, S., Zeng, H., Bai, L., Cai, Z., Zhao, B., et al. (2022). A compendium of genetic regulatory effects across pig tissues. Preprint at bioRxiv. https://doi.org/10.1101/2022.11.11.516073.

24. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods *12*, 931–934. https://doi.org/10.1038/nmeth.3547.

25. Hughes, L.H., Schmitt, M., Mou, L., Wang, Y., and Zhu, X.X. (2018). Identifying Corresponding Patches in SAR and Optical Images With a Pseudo-Siamese CNN. Geosci. Rem. Sens. Lett. IEEE *15*, 784–788. https://doi.org/10.1109/LGRS.2018.2799232.

26. Xiao, S., Xie, D., Cao, X., Yu, P., Xing, X., Chen, C.-C., Musselman, M., Xie, M., West, F.D., Lewin, H.A., et al. (2012). Comparative Epigenomic Annotation of Regulatory DNA. Cell *149*, 1381–1392. https://doi.org/10.1016/j.cell.2012.04.029.

27. Chen, K.M., Wong, A.K., Troyanskaya, O.G., and Zhou, J. (2022). A sequence-based global map of regulatory activity for deciphering human genetics. Nat. Genet. *54*, 940–949. https://doi.org/10.1038/s41588-022-01102-2.

28. Liu, Y., Ma, Y., Yang, J.-Y., Cheng, D., Liu, X., Ma, X., West, F.D., and Wang, H. (2014). Comparative Gene Expression Signature of Pig, Human and Mouse Induced Pluripotent Stem Cell Lines Reveals Insight into Pig Pluripotency Gene Networks. Stem Cell Rev. Rep. *10*, 162–176. https://doi.org/10.1007/s12015-013-9485-9.

29. Beh, L.Y., Colwell, L.J., and Francis, N.J. (2012). A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. Proc. Natl. Acad. Sci. USA *109*, E1063–E1071. https://doi.org/10.1073/pnas.1118678109.

30. Lowy-Gallego, E., Fairley, S., Zheng-Bradley, X., Ruffier, M., Clarke, L., and Flicek, P.; 1000 Genomes Project Consortium (2019). Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. Wellcome Open Res. *4*, 50. https://doi.org/10.12688/wellcomeopenres.15126.2.

31. Flavahan, W.A., Drier, Y., Liau, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suvà, M.L., and Bernstein, B.E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature *529*, 110–114, 110–114. https://doi.org/10.1038/nature16490.

32. Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910. https://doi.org/10.1016/j.cell.2015.07.038.

33. Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer Evolution across 20 Mammalian Species. Cell *160*, 554–566. https://doi.org/10.1016/j.cell.2015.01.006.

34. Yao, Y., Liu, S., Xia, C., Gao, Y., Pan, Z., Canela-Xandri, O., Khamseh, A., Rawlik, K., Wang, S., Li, B., et al. (2022). Comparative transcriptome in large-scale human and cattle populations. Genome Biol. *23*, 176. https://doi.org/10.1186/s13059-022-02745-4.

35. Zhao, R., Talenti, A., Fang, L., Liu, S., Liu, G., Chue Hong, N.P., Tenesa, A., Hassan, M., and Prendergast, J.G.D. (2022). The conservation of human functional variants and their effects across livestock species. Commun. Biol. *5*, 1003. https://doi.org/10.1038/s42003-022-03961-1.

36. Urbut, S.M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. *51*, 187–195. https://doi.org/10.1038/s41588-018-0268-8.

37. Mohammadi, P., Castel, S.E., Brown, A.A., and Lappalainen, T. (2017). Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. Genome Res. *27*, 1872–1884. https://doi.org/10.1101/gr.216747.116.

38. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985. https://doi.org/10.1093/nar/gkt1113.

39. Powell, G., Long, H., Zolkiewski, L., Dumbell, R., Mallon, A.-M., Lindgren, C.M., and Simon, M.M. (2022). Modelling the genetic aetiology of complex disease: human–mouse conservation of noncoding features and disease-associated loci. Biol. Lett. *18*, 20210630. https://doi.org/10.1098/rsbl.2021.0630.

40. Hujoel, M.L.A., Gazal, S., Hormozdiari, F., van de Geijn, B., and 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. https://doi.org/10.1016/j.ajhg.2019.02.008.

41. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235. https://doi.org/10.1038/ng.3404.

42. Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A.P., van de Geijn, B., Reshef, Y., Márquez-Luna, C., et al. (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nat. Genet. *52*, 1355–1363. https://doi.org/10.1038/s41588-020-00735-5.

43. Zheng, Z., Liu, S., Sidorenko, J., Yengo, L., Turley, P., Ani, A., Wang, R., Nolte, I.M., Snieder, H., Yang, J., et al. (2022). Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. Preprint at bioRxiv. https://doi.org/10.1101/2022.10.12.510418.

44. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Series B Stat. Methodol. *82*, 1273–1300. https://doi.org/10.1111/rssb.12388.

45. Zoonomia Consortium; Serres, A., Armstrong, J., Johnson, J., Marinescu, V.D., Murén, E., Juan, D., Bejerano, G., Casewell, N.R., Chemnick, L.G., et al. (2020). A comparative genomics multitool for scientific discovery and conservation. Nature *587*, 240–245. https://doi.org/10.1038/s41586-020-2876-6.

46. Lee, B.T., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, C.M., et al. (2022). The UCSC Genome Browser database: 2022 update. Nucleic Acids Res. *50*, D1115–D1122. https://doi.org/10.1093/nar/gkab959.

47. Chen, K.M., Cofer, E.M., Zhou, J., and Troyanskaya, O.G. (2019). Selene: a PyTorch-based deep learning library for sequence data. Nat. Methods *16*, 315–318. https://doi.org/10.1038/s41592-019-0360-8.

48. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems (Curran Associates, Inc.).

49. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

50. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501. https://doi.org/10.1038/nbt.1630.

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

52. Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human–Mouse Alignments with BLASTZ. Genome Res. *13*, 103–107. https://doi.org/10.1101/gr.809403.

53. 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. https://doi.org/10.1038/s41467-021-22100-8.

54. Pardiñas, A.F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T.B., Bryois, J., Chen, C.-Y., Dennison, C.A., Hall, L.S., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature *604*, 502–508. https://doi.org/10.1038/s41586-022-04434-5.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Human epigenome	ENCODE project[13]	Table S1
Human gene expression	ENCODE project[13]	Table S1
Human chromatin state	Pan et al.[18]	Table S2
Pig epigenome	Pan et al.[18]; Zhao et al.[19]	Table S3
Pig gene expression	Pan et al.[18]; Zhao et al.[19]	Table S6
Pig chromatin state	Pan et al.[18]	Table S4
Orthologous SNPs between human and pig	1,000 Genome Project[30]	http://ftp.1000genomes.ebi.ac.uk
Human GWAS summary statistics	UK Biobank	http://www.ukbiobank.ac.uk
Software and algorithms		
R 4.1	R Core Team	https://www.r-project.org/
Python version 3.6.9	Van Rossum	https://www.python.org/
DeepGCF	This paper	https://github.com/liangend/DeepGCF and https://doi.org/10.5281/zenodo.8087963
LDSC version 1.0.1	Finucane et al.[41]	https://github.com/bulik/ldsc
Polyfun	Weissbrod et al.[42]	https://github.com/omerwe/polyfun
SBayesRC	Zheng et al.[43]	https://github.com/zhilizheng/SBayesRC
R package MashR version 0.2.57	Urbut et al.[36]	https://github.com/stephenslab/mashr
SuSiE	Wang et al.[44]	https://stephenslab.github.io/susieR/index.html
LiftOver	Lee et al.[46]	https://genome.ucsc.edu/cgi-bin/hgLiftOver
Python package selene-sdk	Chen et al.[47]	https://github.com/FunctionLab/selene
Python package torch version 1.10.0	Paszke et al.[48]	https://pypi.org/project/torch/
Python package sklearn version 0.24.2	Pedregosa et al.[49]	https://scikit-learn.org/stable/
R package GREAT version 1.26.0	McLean et al.[50]	https://www.bioconductor.org/packages/release/bioc/html/rGREAT.html
BEDtools version 2.29.1	Quinlan and Hall[51]	http://bedtools.readthedocs.io/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Hao Cheng (qtlcheng@ucdavis.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The DeepGCF scores of humans and pigs, and original codes are available at GitHub: https://github.com/liangend/DeepGCF. The version used in the preparation of the manuscript has been deposited at Zenodo: https://doi.org/10.5281/zenodo.8087963. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Genome alignment

We used the chained and netted alignments of human (GRCh38) and pig (susScr11) genome assemblies from the UCSC genome browser.[46] The assemblies were aligned by the lastz alignment program[52] using human as the reference.

Model inputs

We divided the whole-genome alignment between human and pig into non-overlapping 50-bp regions within each alignment block, resulting in 38,961,848 orthologous pairs. If an alignment block ended shorter than a 50-bp window, the window was truncated to the end of the block, which resulted in some regions smaller than 50 bp. For each orthologous pair, we collected the corresponding functional features, including chromatin accessibility measured by Assay for Transposase-Accessible Chromatin (ATAC-seq), histone modifications measured by Chromatin Immunoprecipitation sequencing (ChIP-seq), chromatin state annotations (ChromHMM), and gene expression measured by RNA-seq for human and pig from public resources, including ENCODE[13] and public literatures.[18,19] We only collected the functional data at the tissue level in humans to make the number of functional features comparable between pigs and humans. We merged binary functional data of the same type from the same tissue into one feature to reduce the computational load. For human, there were 604 ChIP-seq and ATAC-seq files merged into 129 features, 12 ChromHMM files of 15 chromatin states (12 × 15 = 180 features), and 77 RNA-Seq features, which resulted in 386 functional annotations. For pig, there were 287 ChIP-Seq and ATAC-Seq files merged into 84 features, 14 ChromHMM files of 15 chromatin states (14 × 15 = 210 features), and 80 RNA-seq features, which resulted in 374 functional annotations. Details of each data type are reported in Tables S1–S6.

Prediction of binary functional features based on DeepSEA

We trained two DeepSEA models to predict the binary functional features, including ATAC-Seq, ChIP-Seq and chromatin state annotations, of human and pig using the Selene package in Python.[47] We used the peak calls of ATAC-seq and ChIP-seq, and one-hot encoded chromatin state annotations as the training input. We then trained the model based on a sequence region of 1,000 bp, and the feature must take up 50% of the center bin (200 bp) for it to be considered a feature annotated to that sequence. All the hyper-parameters were set as default (Table S19). We created a validation set using the data from chromosomes 6 and 7 for early stopping during training, a test set using the data from chromosomes 8 and 9 for the generation of the receiver operating characteristic (ROC) and precision-recall (PR) curves, and a training set using the rest data. We then predicted the probability of each binary feature using the trained model for the first base of all the paired regions that were at most 50 bp.

Data subsets for training and evaluation

We divided the entire data into the training, validation, testing, and prediction sets based on the chromosome number. To predict the DeepGCF score of human regions from even and X chromosomes and the corresponding paired pig regions (prediction set), we trained a DeepGCF model based on paired regions from a subset of odd chromosomes of human and pig. We created a validation set from another subset of odd chromosomes (not overlapping with the training set) for the hyper-parameter tuning and early stopping during training. A testing set based on paired regions from even chromosomes was used to generate the ROC and PR curves. To predict the DeepGCF score of human regions from odd chromosomes and the corresponding paired pig regions, we created training and validation sets similarly as above, except from even chromosomes, and a testing set from odd chromosomes. We excluded Y and mitochondrial chromosomes in this study. Detailed division of each set is shown in Table S20.

DeepGCF training

Before training the DeepGCF model, we first randomly paired up the human-pig orthologous regions to get an equal number of non-orthologous pairs in the training set. We then trained the DeepGCF model with a pseudo-Siamese architecture similar to the LECIF model.[21] In the pseudo-Siamese neural network, for each orthologous/non-orthologous pair, two input vectors containing the human and pig binary features (probabilities between 0 and 1) predicted from DeepSEA and normalized RNA-seq data (also between 0 and 1) were connected to the human and pig subnetworks, respectively (Figure 1). We performed a natural logarithm transformation on RNA-seq data before normalizing. The two subnetworks were then fully connected to a final subnetwork, which generated the output prediction. We weighted non-orthologous pairs 50 times more than orthologous ones during the training process.

We then used Python packages torch and sklearn to train the DeepGCF model.[48,49] We conducted a random grid search for hyper-parameters, including number of layers in each subnetwork and the final subnetwork, number of neurons in each layer, learning rate, batch size, and dropout rate. We generated 100 combinations of hyper-parameters randomly selected from the candidate parameter pool (Table S21), using each combination to train a DeepGCF model based on the same random subset of 1 million aligned and 1 million unaligned human-pig pairs from the training set. We then selected the combination of hyper-parameters that maximized the AUROC on the validation set to train the final model based on the whole training set. Model training was stopped if there was no improvement in AUROC over three epochs, otherwise it was stopped when reaching the maximum number of epochs, which was set to be 100.

Human-pig orthologous SNPs

In total 73,257,633 human biallelic SNPs (GRCh38) were obtained from 1,000 Genome Project.[30] Their positions were lifted to corresponding orthologous positions in the pig genomes (SusScr11) using the UCSC LiftOver utility,[46] which resulted in 35,575,835 orthologous SNPs.

Function enrichment

To explore the Gene Ontology terms of genomic regions (e.g., enhancers), we used the GREAT tool[50] from with default parameters and a cut-off of FDR <0.05 for both the binomial and the hypergeometric distribution-based tests.

Tissue specific chromatin state

To investigate the tissue specificity of strongly active enhancer and promoter in humans and pigs, we followed the same procedure as described in Pan et al. and Kern et al.[18,53] For each chromatin state, we first used the *merge* function of BEDtools (version 2.29.1)[51] to merge any regulatory regions between two tissues overlapped by 1 bp and obtained a regulatory reference across all tissues. We then used the *intersect* function of BEDtools to find the overlap between regions in the regulatory reference and regulatory file of each tissue. If a region in the reference overlaps with regions in only one tissue, we define the region as tissue-specific regulatory element. If a region overlaps across all tissues, we define the region as "all common" regulatory element.

Tissue-sharing of e/sGene

To explore how e/sGenes (genes with significant e/sQTLs) are shared across all tissues, we performed the meta-analysis of e/sGenes using MashR (v0.2.57).[36] We used the slope and the standard error of slope of top e/sQTL of genes (missing slopes were set to be 0 with standard error of 1) across 49 tissues from GTEx (v8)[22] for human and 34 tissues from PigGTEx databases[23] for pig as the input. We then obtained the estimate of effect size and the corresponding significance (local false sign rate, LFSR) from the mash function. An e/sGene was considered active in a tissue if LFSR <0.05.

DeepGCF score for genes

We obtained the gene boundaries of human and pig genes from Ensembl release 107 (GRCh38 for human and Sscrofa11 for pig), and extended them by 35 kb upstream and 10 kb downstream to include probable *cis*-regulatory regions.[54] We then compute the DeepGCF score for genes based on the average score of all orthologous regions overlapping with the gene and the extended regions. For human genes linked to promoter sequence class, we identified a promoter's potential target gene if the distance between the promoter and the TSS of a gene is less than 2 kb, yielding a total of 12,044 promoter-gene pairs.

Heritability partitioning analysis

We collected the GWAS summary statistics of 80 human complex traits from the UK Biobank and public literatures (Table S14). We ran the LD-score regression software LDSC[41] to partition the heritability based on two sets of annotations: 1) one binary annotation of functionally conserved regions (top 5% of DeepGCF) and 2) five binary annotations dividing the top 5% DeepGCF into 5 equal-width bins based on percentiles. Both sets of annotations were analyzed with a baseline including 97 annotations.[40] Heritability enrichment was calculated as the proportion of trait heritability contributed by SNPs in the annotation over the proportion of SNPs in that annotation.

Fine-mapping analysis

We first used PolyFun[42] to compute SNP prior causal probabilities based on the annotation of functional conservation (top 5% DeepGCF). These probabilities were then used as priors in SuSiE[44] for the fine-mapping analysis. To compare to fine-mapping without using functional conservation as a prior, we also performed a fine-mapping analysis using SuSiE alone, which only took LD information into account. An SNP is identified to be putative causal if the posterior causal probability (PIP) is greater than 0.95 and the p value in GWAS is smaller than 5×10^{-8}.

Polygenic score prediction

We incorporated functional conservation as a prior in polygenic prediction using the software SBayesRC.[43] The GWAS summary statistics of 20 complex traits from UK Biobank (Table S17) were analyzed using ~7 million common SNPs. To compare the prediction accuracy, we partitioned the total sample into ten equal-sized disjoint subsamples. For each fold, we retained one subsample as the validation set and other remaining nine subsamples as the training set. We calculated the polygenic score using genotypes from an independent validation set in each fold and obtained the prediction R^2 from linear regression of the true phenotype on the polygenic score. We then calculated the relative prediction accuracy by $(R_0^2 - R_D^2)/R_0^2$, where R_0^2 is the prediction R^2 without any priors, and R_D^2 is the prediction R^2 using functional conservation as a prior.

QUANTIFICATION AND STATISTICAL ANALYSIS

The quantitative and statistical analyses are described in the relevant sections of the method details.