Quantifying the functional conservation between human and pig using artificial neural networks

J. Li[1], T. Zhao[1], Z. Pan[1], H. Zhou[1], L. Fang[2] and H. Cheng[1*]

[1]Department of Animal Science, University of California, Davis, One Shield Avenue, 95616, Davis, CA, USA; [2]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XU, UK; [*]qtlcheng@ucdavis.edu

Abstract
Identifying functional conservation between human and pig genomes is an important challenge in pig model studies. To evaluate it, we predicted conservation scores through fitting a neural network model based on the integration of multi-tissue epigenetic and gene expression profiles. We then integrated the conservation score with chromatin state and human expression quantitative trait loci (eQTL). The average conservation score for promoters was higher than enhancers as expected. The conservation scores for human regions with eQTL were higher than the average level of the whole genome. In addition, eQTLs with higher conservative scores had a smaller effect size than those with lower scores. In conclusion, the model reflected the genome-wide functional conservation between human and pig, which can be easily extended to other species, and could be further utilized to reveal the conserved patterns linked to complex traits.

Introduction
The investigation of genome conservation not only reveals evolutionary evidence (Alföldi and Lindblad-Toh, 2013), but also helps with the integration of genetic findings across different species (Raymond *et al.*, 2020). The genomic conservation between human and domestic pig (*Sus scrofa*), a relevant biomedical model for certain human diseases, is of importance and interest. Previous studies investigated the functional conservation between human and pig based solely on the gene expression profile (Sjöstedt *et al.*, 2020) or the epigenomic profile (Pan *et al.*, 2021), but not all available functional data were integrated simultaneously to quantify the conservation. Kwon and Ernst (2021) developed a neural network model to study human-mouse functional conservation based on multi-omics information and showed that their method can be used to explain the phenotype-associated variation. In this study, we aim to implement a neural network model to evaluate genome-wide functional conservation between pig and human based on the integration of large-scale epigenetic data and gene expression data across multiple tissues.

Material and methods
The overview of the model development is shown in Figure 1. To define pairs of human and pig regions for training and prediction, we first obtained the pairwise sequence alignment between the human genome (hg38) and the pig genome (Sscrofa11) from the UCSC Genome Browser (Navarro Gonzalez *et al.*, 2021). We then divided the alignment into non-overlapping 50-bp windows, which resulted in 38,961,932 aligned pairs. We randomly paired up human and pig regions included in the aligned pairs to get the same number of unaligned pairs.
For each aligned and unaligned pair, we obtained the corresponding human and pig functional features, including gene expression measured by RNA-seq, chromatin accessibility measured by Assay for Transposase-Accessible Chromatin (ATAC-seq), histone modifications measured by Chromatin Immunoprecipitation sequencing (ChIP-seq), and chromatin state annotations

(ChromHMM). Human features were collected from ENCODE (The ENCODE Project Consortium, 2012) and Roadmap Epigenomics Project (Kundaje *et al.*, 2015), and pig features were from https://doi.org/10.6084/m9.figshare.13480425 (Pan *et al.*, 2021). In total, we collected 313 and 312 functional features for human and pig, respectively.

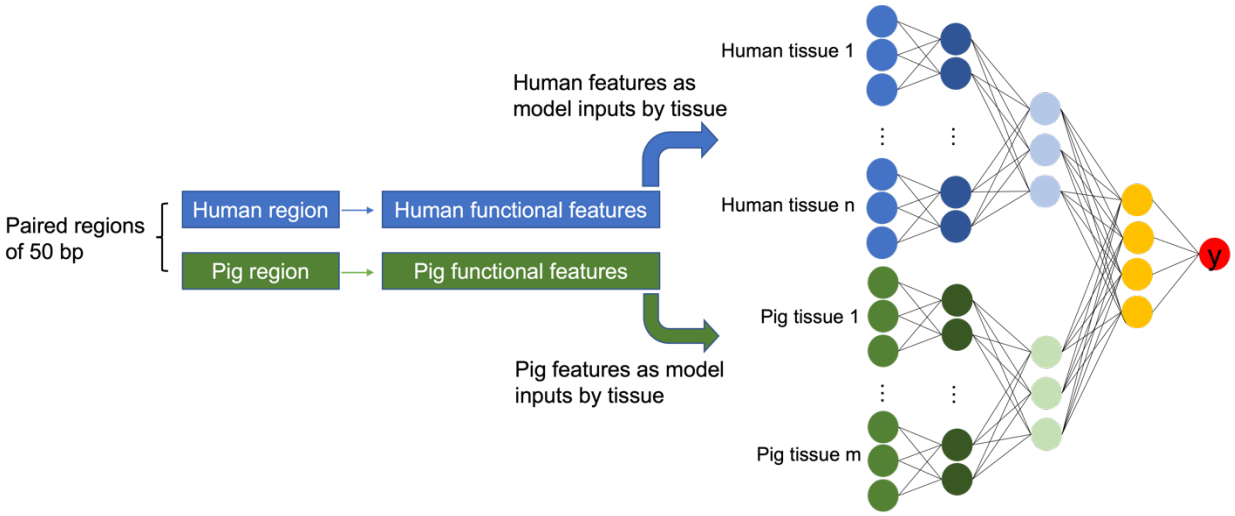

Figure 1. The schematic of the artificial neural network model. The functional features of human and pig in the paired (aligned or unaligned) regions are used as the model inputs by species and tissue, to predict the conservation score (y). When training the model, the response variables (labels) for aligned pairs are coded 1 while those for unaligned pairs are coded 0.

Aligned regions are conservative at the sequence level but not necessarily at the functional level. However, unaligned regions are much less likely to be conservative given different sequence information. Based on this, we presumed the aligned regions to be conservative (coded as 1) and unaligned ones to be unconservative (coded as 0). Thereby a pseudo-Siamese neural network model was trained based on the alignment (1 or 0) as the response variable and corresponding functional features as the predictor variables (Figure 1). To reflect the tissue-specific information in the model structure, the input and first two hidden layers were not fully connected. Instead, the input layer and the first hidden layer were connected by species and tissue, and the first and second hidden layers were connected by species. Although the response variable is binary when training the model, the output was transformed through a sigmoid activation function. Therefore, a continuous conservation score between 0 (not conserved) and 1 (highly conserved), could be predicted for any paired regions based on the functional features.

To predict the conservation score of human regions from even (or odd) chromosomes, and the corresponding paired pig regions, a neural network model was trained based on human odd (or even) chromosomes and paired pig regions. The area under the receiver operating characteristic (ROC) curve was assessed by comparing the predicted results to whether the pairs were aligned or not.

To demonstrate the potential application of the conservative score, we integrated the score with the expression quantitative trait loci (eQTL) in human. The significant loci associated with at least one gene expression from 49 tissues were obtained from the GTEx portal (https://gtexportal.org/home/datasets). We also compared the conservation score and the effect size of eQTLs, which was quantified as log allelic fold change (Mohammadi *et al.*, 2017).

Results and Discussion

The predicted conservation scores showed that most of the aligned pairs (~95%) had a small conservation score (< 0.2), while only ~0.4% of the aligned pairs had a conservation score greater than 0.8. This indicates that only a small part of the aligned genome between human and pig is functionally conserved. The area under the ROC curve was 0.77, suggesting a decent model prediction.

We next investigated average conservation score for chromatin states in each tissue (Figure 2). Similar results were found for human and pig. The highest average conservation score was found in the strongly active promoters, which was higher than all the scores derived from the enhancers. This result was consistent with the previous finding that promoters are more conservative than enhancers in general (Villar *et al.*, 2015).

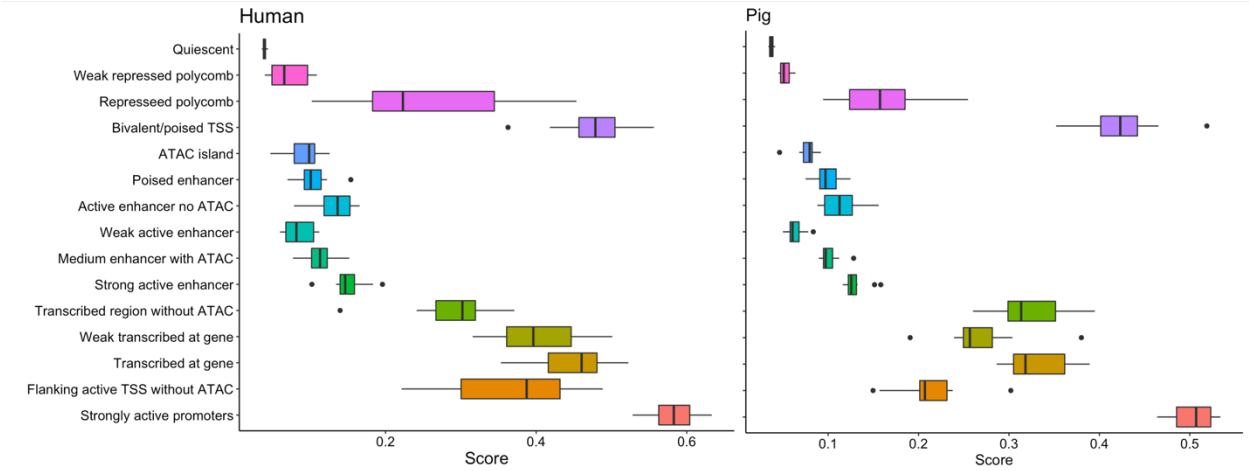

Figure 2. Average conservation scores for different chromatin states across tissues of human (12 tissues) and pig (14 tissues).

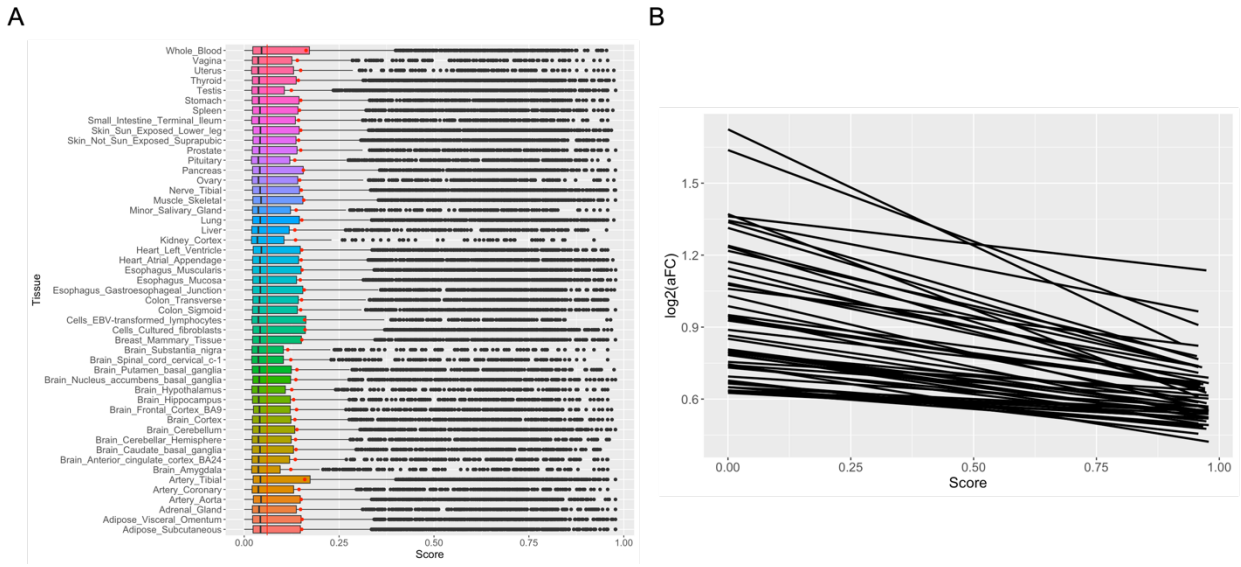

Figure 3. Relationship of conservation scores to human eQTLs. A: Conservation scores on eQTLs across 49 tissues. The red line represents the average score over the whole genome,

and the red dot represents the average score of the eQTLs in the tissue; B: Regression lines of the conservation score and eQTL effect size (log_2(aFC) = log allelic fold change) for 49 tissues.

The results of the application of the conservative score on human eQTLs are shown in Figure 3. The scores obtained by regions with eQTLs were higher than the average score of the whole genome in all tissues (Figure 3A), which suggests that eQTLs regions were more conservative than genomic regions outside eQTLs. We regressed the eQTL effect size on the conservation score and found that the slope was significantly smaller than 0 for all tissues except uterus and blood (Figure 3B). The negative association between effect size and conservation of eQTLs agrees with the previous finding that genomic regions having large effect sizes are less likely to be conservative (Mohammadi *et al.*, 2017).
In conclusion, functional conservation between human and pig predicted by the neural network model had a decent accuracy, and the score reflected the conservation of different chromatin states and eQTLs. The method could be further used to reveal the conserved patterns linked to respective complex traits and diseases between human and pigs, facilitating the recycling information among species. The method can also be easily extended to other species.

References

Alföldi, J., and Lindblad-Toh, K. (2013) Genome Res. 23(7):1063-1068. https://doi.org/10.1101/gr.157503.113

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A. et al. (2015) Nature 518(7539):317-330. https://doi.org/10.1038/nature14248

Kwon, S.B., and Ernst, J. (2021) Nat. Commun. 12:2495. https://doi.org/10.1038/s41467-021-22653-8

Mohammadi, P., Castel, S.E., Brown, A.A., and Lappalainen, T. (2017) Genome Res. 27(11):1872-1884. https://doi.org/10.1101/gr.216747.116

Navarro Gonzalez, J., Zweig, A.S., Speir, M.L., Schmelter, D., Rosenbloom, K.R. et al. (2021) Nucleic Acids Res. 49(D1):D1046-D1057. https://doi.org/10.1093/nar/gkaa1070

Pan, Z., Yao, Y., Yin, H., Cai, Z., Wang, Y. et al. (2021) Nat. Commun. 12:5848. https://doi.org/10.1038/s41467-021-26153-7

Raymond, B., Yengo, L., Costilla, R., Schrooten, C., Bouwman, A.C. et al. (2020) PLoS Genet. 16(9):e1008780. https://doi.org/10.1371/journal.pgen.1008780

Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N. et al. (2020) Science 367(6482):eaay5947. https://doi.org/10.1126/science.aay5947

The ENCODE Project Consortium. (2012) Nature 489(7414):57-74. https://doi.org/10.1038/nature11247

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M. et al. (2015) Cell 160(3):554-566. https://doi.org/10.1016/j.cell.2015.01.006