Title page:

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

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]*, Hao Cheng[1]*

[1]Department of Animal Science, University of California, Davis, Davis, CA, USA

[2]Center for Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, Denmark

[3]Guangdong Laboratory of Lingnan Modern Agriculture, 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, China

[4]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia

[5]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA

*Corresponding authors:

HC: Department of Animal Science, University of California, Davis, Davis, CA, USA

Email: qtlcheng@ucdavis.edu

LF: Center for Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, Denmark

Email: lingzhao.fang@qgg.au.dk

Abstract

The assessment of genomic conservation between human and pig at the functional level can help understand and improve the potential of pig as a human biomedical model. To address this, we developed a **Deep** learning-based approach to learn the **G**enomic **C**onservation at the **F**unctional level (DeepGCF) between species by integrating 386 and 374 epigenome and transcriptome profiles from human and pig, respectively. DeepGCF demonstrated a better prediction performance compared to the previous functional conservation prediction method. In addition, we showed that the resulting DeepGCF score captures the functional conservation by examining DeepGCF on chromatin states, sequence ontologies, and regulatory variants. Regions with higher DeepGCF score play a more important role in regulatory activities and show heritability enrichment in human complex traits and diseases. Our DeepGCF approach shows a promising application on the comparison of cross-species functional conservation, and the model framework can be easily adapted to other species. By expanding the model to integrate the functional profiles of multiple species, including human, mouse, pig, cattle, and other livestock animals in the future, the functional conservation information will provide additional insight into the genetic and evolutionary mechanisms behind complex traits and diseases.

Main text

Introduction

Comparative genome not only reveals evolutionary changes at the DNA sequence level[1], but also helps with the translation of genetic and biological findings across species[2]. Compared to lab organisms like mice, pig is more similar to human in anatomy, physiology, and genome[3], thus is widely used as a biomedical model for human medicine and genetic diseases, such as drug tests[4], xenotransplantation[5], Alzheimer's disease[6], breast cancer[7], and diabetes[8]. To fully recognize the substantial potential of pig as a human biomedical model, it is essential to conduct an extensive comparison of pig and human physiology at the molecular level for assessing to what degree that the genetic and biological findings in pig can be extrapolated to human. Several methods have been proposed to infer the conservation at the DNA sequence level, such as Genomic Evolutionary Rate Profiling (GERP)[9] and Phylogenetic *P*-values (PhyloP)[10]. However, the conservation at DNA sequence level is not equivalent to the conservation at functional level[11–13].

The ongoing global efforts on functional annotation of genomes in both humans and livestock, such as the Encyclopedia of DNA Elements[14], Roadmap Epigenomics projects[15], the Functional Annotation of Animal Genomes (FAANG)[16], and Farm animal Genotype-Tissue Expression (FarmGTEx) projects[17], provide an unprecedented opportunity to quantify the genome conservation across species at the functional level. Previous studies often rely on a single functional profile in one tissue/cell type, such as gene expression[18] or epigenome[19,20], to infer the functional conservation of orthologous regions between human and pig. However, integrative analysis of multi-omics is essential for unravelling how biological information encoded in the genome is conserved or diverged across species, as the functional consequence of

genomic variants is often modulated at multiple levels of gene regulation across tissues/cells. Artificial neural networks have been applied in the prediction and integration of multi-omics data, such as histone marks, transcription factors, and gene expression, to investigate transcriptional and biochemical impact of DNA sequences and their conservation across species[21,22]. For instance, Kwon and Ernst[22] developed a neural network model, LECIF, to study human-mouse functional conservation based on multi-omics data from Roadmap and ENCODE databases.

In this study, to systematically evaluate the functional conservation between human and pig, we developed a **Deep** learning-based approach to learn the **G**enomic **C**onservation at the **F**unctional level (DeepGCF) between species. Unlike LECIF using functional genomics data as input, DeepGCF uses both DNA sequences and functional genomics data as input. It thus enables us to predict the impact of sequence mutations on the functional conservation between species. By integrating 386 and 374 epigenome and transcriptome profiles, representing 28 and 21 tissues from human and pig, respectively, DeepGCF captures the conservation of epigenetic features and genes across tissues between human and pig. By further examining expression/splicing quantitative trait loci (e/sQTL) from 54 and 35 tissues in human GTEx[23] and PigGTEx[24], respectively, and genome-wide association studies (GWAS) of 80 complex traits/diseases in human, DeepGCF provides novel insights into the evolutionary mechanisms underlying both molecular phenotype and complex trait variation. The DeepGCF model can be easily expanded to multiple species for extensively understanding the genome evolution at functional genomics level when large-scale functional annotation data is available for many other species in the near feature.

Results

Overview of the DeepGCF model

In general, the training of DeepGCF model consists of two steps (**Fig. 1**). The first step is to transform the binary functional features to continuous values by training a deep convolutional network implemented in DeepSEA[25]. Binary functional feature is a common data type in the functional genomics filed, which represents whether a genomic base overlapped with functional annotations such as peaks or chromatin states derived from ATAC-Seq and ChIP-Seq. By taking both DNA sequences and binary functional features as inputs, DeepSEA predicts the probabilities of each functional feature at a single-nucleotide resolution. In this study, we collected 309 and 294 genome-wide binary functional annotations from human and pig, respectively (**Supplementary data 1–4**). These represented the chromatin accessibility measured by Assay for Transposase-Accessible Chromatin (ATAC-seq), histone modifications measured by Chromatin Immunoprecipitation sequencing (ChIP-seq) and chromatin states from 26 and 21 tissues in human and pig, respectively. We trained the DeepSEA models and predicted the functional effect of each nucleotide in human and pig separately, which were subsequently used as inputs in the DeepGCF for predicting the functional conservation score between these two species. The performance of DeepSEA was evaluated using an independent validation set and showed a strong predictive power in both species (**Supplementary Fig. 1**).

The second step of DeepGCF is to predict the functional conservation score of orthologous regions between human and pig using a supervised deep learning approach, similar to LECIF[22]. 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 paired alignments (i.e., orthologous regions). We then selected the first base to represent the functional annotation of the 50-bp

108 region, because bases within such a narrow region are likely to have similar functions and the
109 computational burden is greatly lightened by doing so[22]. Apart from the predicted functional
110 effects from DeapSEA, we also included the gene expression values from 77 and 80 RNA-seq
111 datasets as functional annotations, representing 11 and 19 tissues in human and pig, respectively
112 (**Supplementary data 5** and **6**). To train the DeepSEA model, we randomly shifted the human-
113 pig orthologous regions to obtain the same number of non-orthologous pairs. Functional
114 conservation is lack of ground truth, thus as an approximation, we presume that the orthologous
115 regions (coded as 1) are more likely to be functionally conserved than non-orthologous regions
116 (coded as 0). We then trained a pseudo-Siamese neural network model[26] using both functional
117 effects predicted from DeepSEA and gene expression as inputs (**Fig. 1a**). We weighted non-
118 orthologous regions 50 times more than orthologous ones when training to highlight regions with
119 strong evidence of functional conservation[22]. The output, DeepGCF score, is a value between 0
120 and 1 quantifying the functional conservation of the paired human-pig region. Furthermore, since
121 the DeepGCF predicts the functional conservation based on the DNA sequence, it allows us to
122 conduct an *in silico* mutagenesis analysis to assess the impact of orthologous variants on the
123 functional conservation between species through investigating the changes of DeepGCF score
124 caused by a mutation (**Fig. 1b**).
125

126 **The evaluation of DeepGCF model**

127 The performance of DeepGCF was evaluated by predicting whether the paired human-pig
128 regions of an independent testing set are orthologous or not. Compared to LECIF, which had the
129 areas under receiver operating characteristic curve (AUROC) and precision-recall curve
130 (AUPRC) of 0.80 and 0.79, respectively, DeepGCF showed a better predictive ability with

131 AUROC and AUPRC of 0.89 and 0.87, respectively (**Figs. 2a, b**). Of note, we normalized the
132 gene expression values with a natural logarithm transformation, which showed a better predictive
133 ability than that without a transformation (**Supplementary Fig. 2**). Among all the 38,961,848
134 orthologous regions between human and pig, only a small percentage (1.2%) exhibited a
135 DeepGCF score greater than 0.8, while more than half with a score less than 0.1 (**Fig. 2c**),
136 consistent with previous findings between human and mice[22]. This result suggests that most of
137 orthologous regions were not functionally conserved between species.
138 To provide suggestions for researchers who are interested in running the DeepGCF model in
139 other species with limited functional annotation data available, we explored different features
140 that may influence the performance of DeepGCF, including sample size and diversity of
141 functional annotations regarding array and tissue/cell type. When training the model, we
142 downsampled both human and pig functional profiles. We found that using ~50% (Human: 192;
143 Pig: 187) and ~10% (Human: 52; Pig: 47) of the functional profiles resulted in similar AUROC
144 (50%: 0.88; 10%: 0.85) and AUPRC (50%: 0.87; 10%: 0.83) values compared to using all the
145 profiles, but using only ~1% (Human: 4; Pig: 4) of the profiles showed substantially lower
146 AUROC (0.69) and AUPRC (0.68) values (**Fig. 2d**). When leaving one type of functional
147 profiles out, the predictive ability of DeepGCF did not change too much (**Fig. 2e**).
148
149 **Relationship between DNA sequence conservation and functional conservation**
150 To fully explore whether DNA sequence conservation indicates functional conservation, we
151 first examined PhyloP scores, which are commonly used to measure the DNA sequence
152 conservation across species[10]. We observed a U-shaped relationship between PhyloP and
153 DeepGCF scores (**Fig. 3a**), demonstrating that both fast-evolving and slow-evolving sequences

154 exhibited a higher functional conservation between species, compared to evolutionary neutral or
155 near-neutral sequences. This agrees with previous findings on comparing individual epigenetic
156 marks and DNA sequence conservation[19,27]. Furthermore, we defined three types of orthologous
157 regions according to their PhyloP and DeepGCF scores to represent the two tails and the bottom
158 of the U curve: 1) regions with both high DeepGCF (> 95th percentile) and PhyloP (> 95th
159 percentile): high D & high P (n = 260,281), 2) those with high DeepGCF (> 95th percentile) but
160 low PhyloP (< 5th percentile): high D & low P (n = 152,557), and 3) those with low DeepGCF (<
161 5th) and medium PhyloP (between 47.5th and 52.5th): low D & med P (n = 95,231). By examining
162 sequence classes, which are predicted regulatory activities of DNA sequences in human genome
163 by a deep learning model, Sei, trained on a compendium of 21,907 epigenome profiles[28], and
164 Gene Ontology (GO) terms, we found that, compared to the whole genome, high D & high P
165 regions were more enriched in promoter, CTCF, and transcription but depleted in enhancer
166 (Binomial test $P < 0.0001$; **Fig. 3b**). Compared to other regions, high D & high P regions showed
167 a higher enrichment in transcription (Binomial test $P < 0.0001$; **Fig. 3b**), and were significantly
168 associated with several RNA-related regulation processes (**Supplementary Data 7**). This
169 indicates the similarities in transcriptional networks between pig and human[18,29]. High D & low
170 P regions were significantly enriched in Polycomb (Binomial test $P < 0.0001$; **Fig. 3b**), in
171 consistency with the fact that some core subunits of Polycomb protein complexes with similar
172 biological functions have shown a weak evolutionary conservation on DNA sequence across
173 species[30]. The low D & med P regions had similar sequence class compositions as the whole
174 genome background except promoter, which was enriched but to a less extent than high D &
175 high P and high D & low P (Binomial test $P < 0.0001$; **Fig. 3b**), and were enriched in fewer GO
176 terms than regions with high DeepGCF (**Supplementary Data 7–9**). In addition, we examined

177 six different sequence ontologies and found that 5' UTR is the most functionally conserved
178 element, followed by start codon, 3' UTR, stop codon, exon, and finally intron. This is consistent
179 between both human and pig (**Fig. 3c**).
180 To investigate the impact of orthologous variants on the functional conservation between
181 species, we examined 35,575,835 human SNPs that are located in orthologous regions between
182 human and pig, which were obtained from the 1,000 Genome Project[31]. We used the DeepGCF
183 model trained based on only predicted probabilities of binary features from DeepSEA (i.e.,
184 leaving RNA-seq out), as the DeepSEA model does not predict for continuous functional
185 features. The new score predicted from DeepGCF without RNA-seq data had a relatively well
186 agreement with the original DeepGCF score with a Pearson's correlation coefficient (PCC) of
187 0.74 (**Supplementary Fig. 3**). To measure the effect of each human SNP on functional
188 conservation, we recomputed the probabilities of binary features for the corresponding
189 orthologous human region due to the SNP mutation and kept the pig probabilities the same, and
190 used the new probabilities to calculate the updated DeepGCF score. The effect on functional
191 conservation is measured by ΔDeepGCF = DeepGCF after SNP mutation – original DeepGCF. By
192 classifying all the orthologous variants into eight categories[28], we found that most of the variants
193 had a limited effect on the functional conservation (**Fig. 3d**). We further grouped them into 40
194 sequence classes[28], and in general, we found that variants in functional features with larger
195 DeepGCF scores showed the stronger effects on the functional conservation between species
196 (**Fig. 3e)**. Promoter and CTCF were more sensitive to variants than other sequence classes. Of
197 note, the average DeepGCF score of CTCF is lower than that of promoter, but it is much more
198 sensitive to genetic mutations regarding the functional conservation, indicating that the genetic
199 disruption of CTCF binding sites (chromatin conformation) may cause strong impacts on

200 functional genome evolution between species by altering the genome topology and consequently
201 the gene expression[32,33].
202
203 **DeepGCF captures the evolutionary characteristics of regulatory elements**
204 To investigate the functional conservation of distinct regulatory elements between pig and
205 human, we first examined the DeepGCF score of 15 chromatin states predicted from 14 pig
206 tissues and 12 human tissues using ChromHMM[19]. We found that strongly active promoters
207 showed the highest DeepGCF scores (i.e., the strongest functional conservation), followed by
208 poised transcription start site (TSS), chromatin states proximal to TSS, enhancers, and finally
209 repressed Polycomb (**Fig. 4a**). This was consistent between human and pig, which agrees with
210 the conservation properties of regulatory elements reported in the previous studies[19,34]. As
211 chromatin states that play important roles in determining the cellular functions may vary among
212 different tissues, we identified strongly active promoters and enhancers that were specific in each
213 of 12 human tissues and 14 pig tissues. Compared to promoters and enhancers shared across all
214 the tissues, tissue-specific ones showed significantly lower DeepGCF scores in both species
215 (Mann–Whitney U test $P < 2.2e-16$), indicative of their faster evolutionary rate (**Fig. 4b**). Among
216 eight common tissues between human and pig, we found that adipose had the strongest
217 functionally conserved promoters in both human and pig, followed by spleen, lung, cortex, liver,
218 and finally stomach (**Supplementary Fig. 4a**). This result suggests pigs could be a good model
219 animal for studying human obesity and metabolic traits[19]. However, the tissue-conservation
220 patterns of enhancers were different from those of promoters and were not consistent between
221 species (**Supplementary Fig. 4b**).

222 We further investigated the DeepGCF score on human promoters and enhancers annotated by
223 Sei[28]. We linked a promoter to its potential target gene and then ranked genes with the DeepGCF
224 scores of their promoters (from largest to smallest). We found that top 5% of genes were
225 significantly enriched in basic biological processes, such as anatomical structure development
226 and organ morphogenesis, whereas bottom 5% of genes were significantly enriched in
227 biosynthetic and metabolic process (**Supplementary data 10 and 11**). In addition, we ranked
228 enhancers according to their own DeepGCF scores and investigated the function of top 5% and
229 bottom 5% enhancers. Unlike promoters, top 5% of enhancers exhibited the most significant
230 enrichment in metabolic processes, while bottom 5% of enhancers were significantly enriched in
231 organ growth and development (**Supplementary data 12 and 13**). In general, we found that
232 promoters and enhancers with a higher DeepGCF score were enriched in much more biological
233 processes compared to those with a lower DeepGCF score (**Fig. 4c, d**), which indicates that
234 functionally conserved regions between species tend to be the hotspot of regulatory activities.
235

236 **DeepGCF provide insight into the functional conservation of regulatory variants**

237 To explore the functional conservation of regulatory variants, we systematically examined
238 expression QTLs (eQTLs) and splicing QTLs (sQTLs) falling in the orthologous regions in 54
239 human tissues and 35 pig tissues, respectively. In general, DeepGCF scores of eQTLs and sQTLs
240 were significantly (Mann–Whitney U test $P < 2.2e-16$) higher than the genome background
241 across all the tissues in both human and pig (**Fig. 5a; Supplementary Figs. 5 and 6**), which
242 suggests that regulatory variants are functionally conserved between species[35,36]. Of note, sQTLs
243 showed a higher DeepGCF score than eQTLs in both species (Mann–Whitney U test $P < 1e-8$),
244 probably due to their larger impacts on the transcriptome function (underlying a stronger

245 purifying selection). This is consistent with previous findings that sQTLs were more likely to be
246 enriched in 5'UTR than eQTLs (GTEx, 2020), and 5' UTR is the most functionally conserved
247 genomic features (**Fig. 2c**). We further observed that eGenes associated with eQTLs having a
248 larger absolute effect on the gene expression had a lower DeepGCF score in both species (**Fig.**
249 **5b**), which suggests that orthologous regions with smaller regulatory effects are more likely to be
250 functionally conserved between species, probably due to the stronger purifying selection
251 underlying them[37]. Moreover, regulatory variants influencing more tissues showed higher
252 DeepGCF scores (i.e., more functionally conserved), consistent in human and pig (**Fig. 5c, d**). In
253 addition, the tissue-sharing pattern of orthologous eGenes (PCC = 0.38, *P* value < 2.2e-16) and
254 sGenes (PCC = 0.45, *P* value < 2.2e-16) were positively correlated between human and pig.
255 Altogether, these results indicate that regulatory variants controlling transcriptome function in
256 more tissues tend to be more functionally conserved between species.
257 We then investigated the DeepGCF scores of 105,461 pathological and likely pathological
258 SNPs obtained from the ClinVar database[38]. A total 98.6% of these SNPs were in the human-pig
259 orthologous regions, consistent with a previous finding that reported more than 98% of
260 pathological variants of Mendelian diseases located in human-mouse orthologous regions[39].
261 Compared to random orthologous regions, these pathological SNPs were significantly more
262 functionally conserved (Mann–Whitney U test *P* < 2.2e-16; **Fig. 6a**). Like orthologous SNP, we
263 classified the ClinVar SNP into eight sequence class categories[28] and conducted an *in silico*
264 mutagenesis analysis to predict their impact on the functional conservation. Overall, the average
265 magnitude of variant effect (measured by |ΔDeepGCF|) for pathological and likely pathological
266 mutations is 1.5 times larger than that for random orthologous SNPs (0.0088 versus 0.0058,
267 Mann–Whitney U test *P* < 2.2e-16). In most of cases, the DeepGCF score did not change much

268 after genetic mutations, but the variance of ΔDeepGCF showed a bell-shaped curve regarding the
269 original DeepGCF score, indicating that SNPs with a medium-high DeepGCF (50th to 80th
270 percentile) were more sensitive to pathological mutations than those with lower or higher
271 DeepGCF (**Fig. 6b**). This suggests that the most functionally conserved regions (> 90th
272 percentile) are more tolerable of mutations than less conserved ones (50th to 80th percentile).
273 Most of the ClinVar SNPs were classified as transcription (51.2%), followed by enhancer
274 (16.4%), Polycomb (14.8%), promoter (8.8%), transcription factor (3.3%), and CTCF (2.2%;
275 **Fig. 6c**). Among the ClinVar SNPs with top 5% of |ΔDeepGCF| (> 0.03), there were more SNPs
276 relevant to a decreased DeepGCF (54.4%) than an increased one (45.6%). Moreover, 9 out of 10
277 ClinVar SNPs with the largest effect on DeepGCF were relevant to a decreased DeepGCF (**Fig.**
278 **6c**). In summary, pathological and likely pathological SNPs are located in functionally more
279 conserved regions, and their impact on functional conservation tends to be related to a decreased
280 functional conservation between human and pig.
281
282 **Application of DeepGCF on gene mapping and prediction for human complex traits**
283 To investigate whether DeepGCF scores could advance our understanding of the evolutionary
284 basis of complex traits/diseases in human, we conducted a heritability partitioning analysis used
285 the functionally conserved genomic regions (top 5% DeepGCF scores) as a functional
286 annotation, along with 97 existing annotations from the baseline model of LDSC[40,41], to analyze
287 the GWAS summary statistics from 80 human complex traits/diseases (**Supplementary Data**
288 **14**). We found that regions with higher DeepGCF scores explained more heritability of complex
289 traits/diseases (**Fig. 7a**). The heritability of eight complex traits was significantly enriched in
290 functionally conserved regions, with the most enrichment found for coxarthrosis (enrichment =

291 3.5, FDR = 0.032), followed by varicose veins, height, hypertension, primary hypertension,
292 waist-hip ratio, weight, and BMI (**Supplementary Data 15; Fig. 7b**). Furthermore, we took
293 these eight traits as examples to explore whether DeepCGF can help us with fine-mapping of
294 causal variants. By using functionally conserved regions (top 5% of DeepCGF) as a biological
295 prior in the PolyFun + SuSiE model[42], we detected 33, 22, and 17 additional putative causal
296 variants (PIP > 0.95 and P < 5e-8) compared to the SuSiE model only without any priors in
297 height, BMI and weight, respectively (**Fig. 7c,** Supplementary Data 16). We further incorporated
298 DeepCGF in SBayesRC[43] model to conduct polygenic score prediction for 20 human complex
299 traits (**Supplementary Data 17**). On average, the relative prediction accuracy increased by
300 0.56% (**Fig. 7d; Supplementary Data 18**), and the largest increase was observed on waist-hip
301 ratio (3.5%), followed by body weight (1.7%). Altogether, our results showed that DeepGCF
302 provide additional insights into the genetic and evolutionary basis of complex phenotypes.
303
304 **Discussion**
305 In this study, we developed a two-step neural network approach, DeepGCF, to evaluate the
306 genomic conservation at the functional level between human and pig. DeepGCF shares a similar
307 model structure as LECIF[22] in the evaluation of functional conservation by comparing the
308 epigenome and gene expression profiles of orthologous regions between two species. But instead
309 of using binary epigenome profiles as the direct inputs, DeepGCF first predicts their functional
310 effects (i.e., the continuous probability score of each epigenome binary feature) using
311 DeepSEA[25], and then use them as the input to predict the functional conservation between
312 species. Compared to the LECIF approach, DeepSEA showed a better performance in the
313 ortholog prediction, probably due to a higher resolution of the model input. Similar to LECIF, we

314 found that the performance of DeepGCF was not sensitive to the number of functional features,
315 indicating that DeepGCF could be applied on other species where functional features are not
316 abundant.

317 We demonstrated that functional conservation is different from sequence conservation. The
318 relationship between DeepGCF and PhyloP scores confirms the U shape relationship between
319 functional and sequence conservation. By examining DeepGCF on chromatin states, sequence
320 ontologies, and regulatory variants, we verified that DeepGCF captures the functional
321 conservation of genome, and regions with higher DeepGCF play a more important role in
322 regulatory activities. We thereby expected DeepGCF to be useful in explaining complex traits
323 and diseases. The heritability enrichment and polygenic prediction accuracy brought by
324 functionally conserved regions were limited, this may because we only considered functional
325 conservation between human and pig compared to sequence conservation which were obtained
326 based on over 100 species[44]. With the increasing amount of epigenome and gene expression data
327 in other species in the near future, we could identify the core functionally conserved regions by
328 expanding the DeepGCF model structure to integrate functional profiles from multiple species.
329 Another limitation is that the functional conservation of the same sequence segment in different
330 tissues and cell types should be conceptually different, which could not be distinguished by the
331 current DeepGCF score. One ideal way to obtain tissue- and cell-type- specific DeepGCF scores
332 is to train a different model on each tissue and cell type using the respective data. However, the
333 current volume of functional profiles, particularly in pig, does not support the development of
334 tissue- and cell-type- specific DeepGCF models.

335 Despite the limitations, the DeepGCF approach shows a promising application on the
336 comparison of cross-species functional conservation. The model framework can be easily

337 adapted to other species. Our future work will focus on expanding the model to the comparison
338 of multiple species, including human, mouse, pig, cattle, and other livestock animals. The
339 functional conservation information among different species will provide additional insight into
340 the genetic and evolutionary mechanisms behind complex traits and diseases, analogous to the
341 sequence conservation among vertebrate animals provided by such as PhyloP score.
342

343 **Methods**

344 **Genome alignment.** We used the chained and netted alignments of human (GRCh38) and pig
345 (susScr11) genome assemblies from the UCSC genome browser[45]. The assemblies were aligned
346 by the lastz alignment program[46] using human as the reference.

347 **Model inputs.** We divided the whole-genome alignment between human and pig into non-
348 overlapping 50-bp regions within each alignment block, resulting in 38,961,848 orthologous
349 pairs. If an alignment block ended shorter than a 50-bp window, the window was truncated to the
350 end of the block, which resulted in some regions smaller than 50 bp. For each orthologous pair,
351 we collected the corresponding functional features, including chromatin accessibility measured
352 by Assay for Transposase-Accessible Chromatin (ATAC-seq), histone modifications measured
353 by Chromatin Immunoprecipitation sequencing (ChIP-seq), chromatin state annotations
354 (ChromHMM), and gene expression measured by RNA-seq for human and pig from public
355 resources, including ENCODE[14] and public literatures[19,20]. We only collected the functional data
356 at the tissue level for human, and merged those of the same data type from the same tissue, so
357 that the total number of human features were close to pig. For human, there were 604 ChIP-seq
358 and ATAC-seq files merged into 129 features, 12 ChromHMM files of 15 chromatin states (12 ×
359 15 = 180 features), and 77 RNA-seq features, which resulted in 386 functional annotations. For

360 pig, there were 287 ChIP-seq and ATAC-seq files merged into 84 features, 14 ChromHMM files
361 of 15 chromatin states (14 × 15 = 210 features), and 80 RNA-seq features, which resulted in 374
362 functional annotations. Details of features from each data type are reported in Supplementary
363 Data 1–6.

364 **Prediction of binary functional features based on DeepSEA.** We trained two DeepSEA
365 models to predict the binary functional features, including ATAC-seq, ChIP-seq and chromatin
366 state annotations, of human and pig using the PyTorch-based package, Selene[47]. We used the
367 peak calls of ATAC-seq and ChIP-seq, and one-hot encoded chromatin state annotations as the
368 training input. We then trained the model based on a sequence region of 1,000 bp, and the feature
369 must take up 50% of the center bin (200 bp) for it to be considered a feature annotated to that
370 sequence. All the hyperparameters were set as default (Supplementary Data 19). We created a
371 validation set using the data from chromosomes 6 and 7 for early stopping during training, a test
372 set using the data from chromosomes 8 and 9 for the generation of the receiver operating
373 characteristic (ROC) and precision-recall (PR) curves, and a training set using the rest data. We
374 then predicted the probability of each binary feature using the trained model for the first base of
375 all the paired regions that were at most 50 bp.

376 **Data subsets for training and evaluation.** We divided the entire data into the training,
377 validation, and prediction sets based on the chromosome number. To predict the DeepGCF score
378 of human regions from even and X chromosomes (prediction set), and the corresponding paired
379 pig regions, we trained a DeepGCF model based on paired regions from a subset of odd
380 chromosomes of human and pig. We created a validation set also from another subset of odd
381 chromosomes (not overlapping with the training set) for the hyper-parameter tuning and early
382 stopping during training. We used a subset of the test set to generate the ROC and PR curves. To

383 predict the DeepGCF score of human regions from odd chromosomes and the corresponding
384 paired pig regions, we created training and validation set similarly as above, except from even
385 chromosomes. We excluded Y and mitochondrial chromosomes in this study. Detailed division
386 of each set is shown in Supplementary Data 20.
387 **DeepGCF training.** Before training the DeepGCF model, we first randomly paired up the
388 human-pig orthologous regions to get the same number of non-orthologous pairs in the training
389 set. We then trained the DeepGCF model with a pseudo-Siamese architecture as the LECIF
390 model[22]. In our pseudo-Siamese neural network, for each orthologous/non-orthologous pair, two
391 input vectors containing the human and pig binary features (probabilities between 0 and 1)
392 predicted from DeepSEA and normalized RNAseq data (also between 0 and 1) were connected
393 to the human and pig subnetworks, respectively (Fig. 1). We performed a natural logarithm
394 transformation on RNAseq data given the large range before normalizing. The two subnetworks
395 were then fully connected to a final subnetwork, which generated the output prediction. We
396 weighted non-orthologous pairs 50 times more than orthologous ones during the training process.
397 We conducted a random grid search for hyper-parameters, including number of layers in each
398 subnetwork and the final subnetwork, number of neurons in each layer, learning rate, batch size,
399 and dropout rate. We generated 100 combinations of hyper-parameters randomly selected from
400 the candidate parameter pool (Supplementary Data 21), using each combination to train a
401 DeepGCF model based on the same random subset of 1 million aligned and 1 million unaligned
402 human-pig pairs from the training set. We then selected the combination of hyper-parameters
403 that maximized the AUROC on the validation set to train the final model based on the whole
404 training set. We stopped training if there was no improvement in AUROC over three epochs on

405 the validation set for both hyper-parameter search and training, otherwise the training stopped
406 when reaching the maximum number of epochs, which was set to be 100.
407 **Human-pig orthologous SNPs.** In total 73,257,633 human biallelic SNPs (GRCh38) were
408 obtained from 1,000 Genome Project[31]. Their positions were lifted to corresponding orthologous
409 positions in the pig genomes (SusScr11) using the UCSC liftover utility with chain files available
410 from the UCSC website[45], which resulted in 35,575,835 orthologous SNPs.
411 **Function enrichment.** To explore the Gene Ontology terms of genomic regions (e.g.,
412 enhancers), we used the GREAT tool[48] with default parameters and a cut-off of FDR < 0.05 for
413 both the binomial and the hypergeometric distribution-based tests.
414 **Tissue specific chromatin state.** For each chromatin state, we first used the merge function of
415 BEDtools[49] to merge any regulatory regions between two tissues overlapping by at least 1 bp
416 across all tissues. Then for strongly active enhancer and promoter in each tissue, if a region is
417 active in only one tissue and does not overlap with any active regions in other tissues, we define
418 the region as tissue specific regulatory element. If a region is active in all tissues (i.e., overlaps
419 across all tissues), we define the region as "all common" regulatory element.
420 **Tissue-sharing of e/sGene.** To explore how e/sGenes are shared across all tissues, we performed
421 the meta-analysis of e/sGenes using MashR (v0.2.57)[50]. We used the slope and the standard error
422 of slope of top e/sQTL of genes (missing slopes were set to be 0 with standard error of 1) across
423 49 tissues from GTEx (v8)[23] for human and 34 tissues from PigGTEx databases[24] for pig as the
424 input. We then obtained the estimate of effect size and the corresponding significance (local false
425 sign rate, LFSR) from the mash function. An e/sGene was considered active in a tissue if LFSR
426 < 0.05.

427 **DeepGCF score for genes.** We obtained the gene boundaries of human and pig genes from
428 Ensembl release 107 (GRCh38 for human and Sscrofa11 for pig), and extended them by 35 kb
429 upstream and 10 kb downstream to include probable cis-regulatory regions[51]. We then compute
430 the DeepGCF score for genes based on the average score of all orthologous regions overlapping
431 with the gene and the extended regions. For human genes linked to promoter sequence class, we
432 identified a promoter's potential target gene if the distance between the promoter and the TSS of
433 a gene is less than 2 kb, yielding a total of 12,044 promoter-gene pairs.
434 **Heritability partitioning analysis.** We collected the GWAS summary statistics of 80 human
435 complex traits from the UK Biobank and public literatures (Supplementary Data 14). We ran the
436 LD-score regression software ldsc (v1.0.1)[41] to partition the heritability based on two sets of
437 annotations: 1) one binary annotation of functionally conserved regions (top 5% of DeepGCF)
438 and 2) five binary annotations dividing the top 5% DeepGCF into 5 equal-width bins based on
439 percentiles. Both sets of annotations were analyzed with a baseline including 97 annotations[40].
440 Heritability enrichment was calculated as the proportion of trait heritability contributed by SNPs
441 in the annotation over the proportion of SNPs in that annotation.
442 **Fine-mapping analysis.** We first used PolyFun[42] to compute SNP prior causal probabilities
443 based on the annotation of functional conservation (top 5% DeepGCF). These prior causal
444 probabilities were then used as priors in SuSiE[52] for the fine-mapping analysis. To compare fine-
445 mapping using functional conservation as prior with not using it, we also performed a fine-
446 mapping analysis using SuSiE alone, which only took LD information into account. A SNP is
447 identified to be putative causal if the posterior causal probability (PIP) is greater than 0.95 and
448 the *P*-value in GWAS is smaller than 5e-8.

449 **Polygenic prediction**. We incorporated functional conservation as a prior in polygenic
450 prediction using the software SBayesRC[43]. The GWAS summary statistics of 20 complex traits
451 from UK Biobank (Supplementary Data 17) were analyzed using ~7 million common SNPs with
452 and without one annotation of functional conservation (top 5% DeepGCF). To compare the
453 prediction accuracy, we partitioned the total sample into ten equal-sized disjoint subsamples. For
454 each fold, we retained one subsample as the validation set and other remaining nine subsamples
455 as the training set. We calculated the polygenic score (PGS) using genotypes from an
456 independent validation set in each fold and obtained the prediction R^2 from linear regression of
457 true phenotype on the PGS. We then calculated the relative prediction accuracy by $(R_0^2 – R_D^2)$ /
458 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
459 functional conservation as a prior.
460
461 **Data availability**
462 The DeepGCF score for human-pig orthologous regions are publicly available for download
463 without restrictions from https://github.com/liangend/DeepGCF. All epigenomic and gene
464 expression data used for model training can be found in Supplementary data 1–6. Orthologous
465 SNPs between human and pig are from the 1,000 Genome Project
466 (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/2018
467 1203_biallelic_SNV). GWAS summary statistics used for LDSC analysis are from UK Biobank
468 (http://www.ukbiobank.ac.uk), with details showing in Supplementary data 14. Summary
469 statistics and genotype used for polygenic score prediction from UK Biobank
470 (http://www.ukbiobank.ac.uk) are available through formal application.
471

472 **Code availability**

473 The code of DeepGCF is available at https://github.com/liangend/DeepGCF.

474

475 **References**

476 1. Alföldi, J. & Lindblad-Toh, K. Comparative genomics as a tool to understand evolution and
477 disease. *Genome Res.* **23**, 1063–1068 (2013).
478 2. Raymond, B. *et al.* Using prior information from humans to prioritize genes and gene-
479 associated variants for complex traits in livestock. *PLOS Genetics* **16**, e1008780 (2020).
480 3. Lunney, J. K. *et al.* Importance of the pig as a human biomedical model. *Science*
481 *Translational Medicine* **13**, eabd5758 (2021).
482 4. Schelstraete, W., Devreese, M. & Croubels, S. Comparative toxicokinetics of Fusarium
483 mycotoxins in pigs and humans. *Food and Chemical Toxicology* **137**, 111140 (2020).
484 5. Montgomery, R. A. *et al.* Results of Two Cases of Pig-to-Human Kidney
485 Xenotransplantation. *New England Journal of Medicine* **386**, 1889–1898 (2022).
486 6. Kragh, P. M. *et al.* Hemizygous minipigs produced by random gene insertion and handmade
487 cloning express the Alzheimer's disease-causing dominant mutation APPsw. *Transgenic Res*
488 **18**, 545–558 (2009).
489 7. Luo, Y. *et al.* High efficiency of BRCA1 knockout using rAAV-mediated gene targeting:
490 developing a pig model for breast cancer. *Transgenic Res* **20**, 975–988 (2011).
491 8. Renner, S. *et al.* Permanent Neonatal Diabetes in INSC94Y Transgenic Pigs. *Diabetes* **62**,
492 1505–1511 (2013).
493 9. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic
494 sequence. *Genome Res.* **15**, 901–913 (2005).

495 10. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral
496 substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
497 11. Bordeira-Carriço, R. *et al.* Multidimensional chromatin profiling of zebrafish pancreas to
498 uncover and investigate disease-relevant enhancers. *Nat Commun* **13**, 1945 (2022).
499 12. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human
500 embryonic stem cells. *Nat Genet* **42**, 631–634 (2010).
501 13. Pennacchio, L. A. & Visel, A. Limits of sequence and functional conservation. *Nat Genet* **42**,
502 557–558 (2010).
503 14. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project.
504 *Science* **306**, 636–640 (2004).
505 15. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**,
506 317–330 (2015).
507 16. Andersson, L. *et al.* Coordinated international action to accelerate genome-to-phenome with
508 FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology* **16**, 57
509 (2015).
510 17. Liu, S. *et al.* A multi-tissue atlas of regulatory variants in cattle. *Nat Genet* 1–10 (2022)
511 doi:10.1038/s41588-022-01153-5.
512 18. Sjöstedt, E. *et al.* An atlas of the protein-coding genes in the human, pig, and mouse brain.
513 *Science* **367**, eaay5947 (2020).
514 19. Pan, Z. *et al.* Pig genome functional annotation enhances the biological interpretation of
515 complex traits and human disease. *Nat Commun* **12**, 5848 (2021).
516 20. Zhao, Y. *et al.* A compendium and comparative epigenomics analysis of cis-regulatory
517 elements in the pig genome. *Nat Commun* **12**, 2217 (2021).

518 21. Wong, A. K., Sealfon, R. S. G., Theesfeld, C. L. & Troyanskaya, O. G. Decoding disease:
519 from genomes to networks to phenotypes. *Nat Rev Genet* **22**, 774–790 (2021).
520 22. Kwon, S. B. & Ernst, J. Learning a genome-wide score of human–mouse conservation at the
521 functional genomics level. *Nat Commun* **12**, 2495 (2021).
522 23. THE GTEX CONSORTIUM *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis:
523 Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
524 24. Consortium, T. F.-P. *et al.* A compendium of genetic regulatory effects across pig tissues.
525 2022.11.11.516073 Preprint at https://doi.org/10.1101/2022.11.11.516073 (2022).
526 25. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–
527 based sequence model. *Nat Methods* **12**, 931–934 (2015).
528 26. Hughes, L. H., Schmitt, M., Mou, L., Wang, Y. & Zhu, X. X. Identifying Corresponding
529 Patches in SAR and Optical Images With a Pseudo-Siamese CNN. *IEEE Geoscience and*
530 *Remote Sensing Letters* **15**, 784–788 (2018).
531 27. Xiao, S. *et al.* Comparative Epigenomic Annotation of Regulatory DNA. *Cell* **149**, 1381–
532 1392 (2012).
533 28. Chen, K. M., Wong, A. K., Troyanskaya, O. G. & Zhou, J. A sequence-based global map of
534 regulatory activity for deciphering human genetics. *Nat Genet* **54**, 940–949 (2022).
535 29. Liu, Y. *et al.* Comparative Gene Expression Signature of Pig, Human and Mouse Induced
536 Pluripotent Stem Cell Lines Reveals Insight into Pig Pluripotency Gene Networks. *Stem Cell*
537 *Rev and Rep* **10**, 162–176 (2014).
538 30. Beh, L. Y., Colwell, L. J. & Francis, N. J. A core subunit of Polycomb repressive complex 1
539 is broadly conserved in function but not primary sequence. *Proceedings of the National*
540 *Academy of Sciences* **109**, E1063–E1071 (2012).

541 31. Lowy-Gallego, E. *et al.* Variant calling on the GRCh38 assembly with the data from phase
542 three of the 1000 Genomes Project. *Wellcome Open Res* **4**, 50 (2019).
543 32. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas.
544 *Nature* **529**, 110–114 (2016).
545 33. Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and
546 Enhancer/Promoter Function. *Cell* **162**, 900–910 (2015).
547 34. Villar, D. *et al.* Enhancer Evolution across 20 Mammalian Species. *Cell* **160**, 554–566
548 (2015).
549 35. Yao, Y. *et al.* Comparative transcriptome in large-scale human and cattle populations.
550 *Genome Biology* **23**, 176 (2022).
551 36. Zhao, R. *et al.* The conservation of human functional variants and their effects across
552 livestock species. *Commun Biol* **5**, 1–13 (2022).
553 37. Mohammadi, P., Castel, S. E., Brown, A. A. & Lappalainen, T. Quantifying the regulatory
554 effect size of cis-acting genetic variation using allelic fold change. *Genome Res.* **27**, 1872–
555 1884 (2017).
556 38. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and
557 human phenotype. *Nucleic Acids Research* **42**, D980–D985 (2014).
558 39. Powell, G. *et al.* Modelling the genetic aetiology of complex disease: human–mouse
559 conservation of noncoding features and disease-associated loci. *Biology Letters* **18**,
560 20210630.
561 40. Hujoel, M. L. A., Gazal, S., Hormozdiari, F., van de Geijn, B. & Price, A. L. Disease
562 Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient

563 Sequence Age and Conserved Function across Species. *The American Journal of Human*
564 *Genetics* **104**, 611–624 (2019).
565 41. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide
566 association summary statistics. *Nat Genet* **47**, 1228–1235 (2015).
567 42. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of
568 complex trait heritability. *Nat Genet* **52**, 1355–1363 (2020).
569 43. Zheng, Z. *et al.* Leveraging functional genomic annotations and genome coverage to improve
570 polygenic prediction of complex traits within and between ancestries. 2022.10.12.510418
571 Preprint at https://doi.org/10.1101/2022.10.12.510418 (2022).
572 44. Genereux, D. P. *et al.* A comparative genomics multitool for scientific discovery and
573 conservation. *Nature* **587**, 240–245 (2020).
574 45. Lee, B. T. *et al.* The UCSC Genome Browser database: 2022 update. *Nucleic Acids Research*
575 **50**, D1115–D1122 (2022).
576 46. Schwartz, S. *et al.* Human–Mouse Alignments with BLASTZ. *Genome Res.* **13**, 103–107
577 (2003).
578 47. Chen, K. M., Cofer, E. M., Zhou, J. & Troyanskaya, O. G. Selene: a PyTorch-based deep
579 learning library for sequence data. *Nat Methods* **16**, 315–318 (2019).
580 48. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions.
581 *Nat Biotechnol* **28**, 495–501 (2010).
582 49. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
583 features. *Bioinformatics* **26**, 841–842 (2010).

50. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* **51**, 187–195 (2019).

51. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).

52. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1273–1300 (2020).

Figures and legends

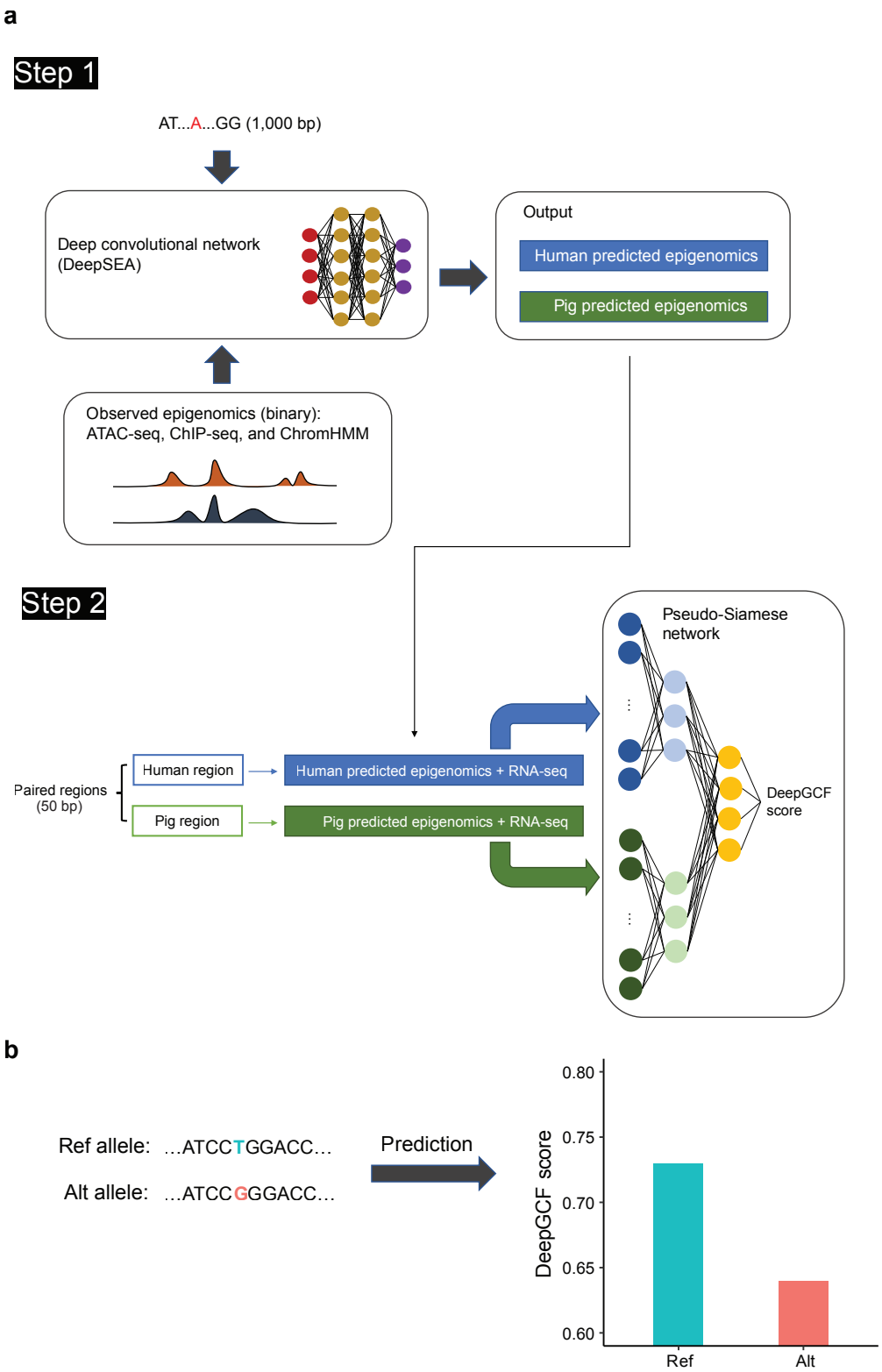

Fig. 1 Overview of the DeepGCF model. a The learning procedure of DeepGCF model consists of two steps. The first step is to train DeepSEA models in human and pig separately to transform the binary functional features (e.g., peaks called from ATAC-seq and ChIP-seq, and chromatin states predicted from 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 for predicting whether the paired human-pig regions are orthologous or not using two corresponding vectors of functional effects predicted from DeepSEA and normalized gene expression as inputs. 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 the functional conservation, quantified by changes in DeepGCF scores.

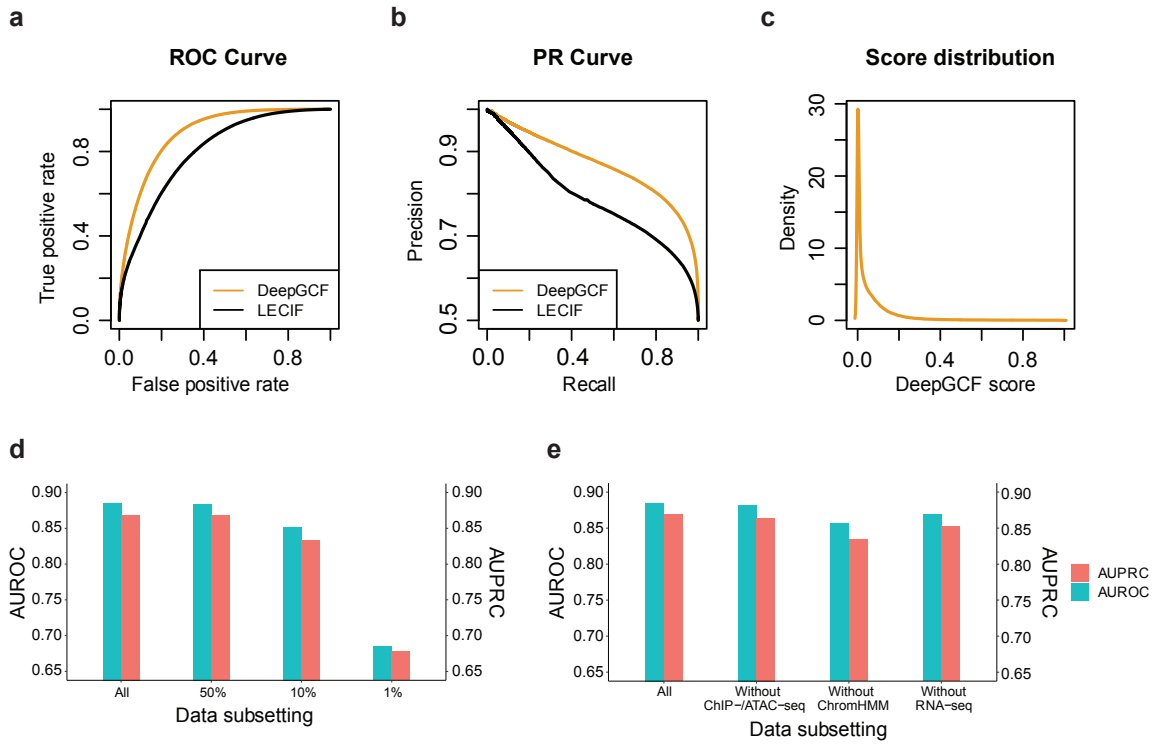

Fig. 2 The performance of DeepGCF under different scenarios a Receiver operating characteristic (ROC) curves comparing the performance of DeepGCF (this study) and LECIF[22] methods. The ROC curve of each method is generated by predicting whether 200,000 pairs randomly selected from the testing set, which included equal number of orthologous and non-orthologous pairs (e.g., randomly mismatched genomics regions), were orthologous or not. **b** Precision-recall (PR) curves generated by similar procedures as the ROC curves. **c** DeepGCF score distribution of all 38,961,848 human-pig orthologues pairs. **d** The areas under receiver operating characteristic curve (AUROC) and precision-recall 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 selected ~50%, ~10%, ~1% from each of ChIP-/ ATAC-seq, ChromHMM, and RNAseq profiles. **e** The AUROC and AUPRC of DeepGCF using all

620 functional features (Human: 386; Pig: 374), features without ChIP-/ATAC-seq (Human: 129;
621 Pig: 84), without ChromHMM (Human: 180; Pig: 210) and without RNA-seq (Human: 77; Pig:
622 80).
623

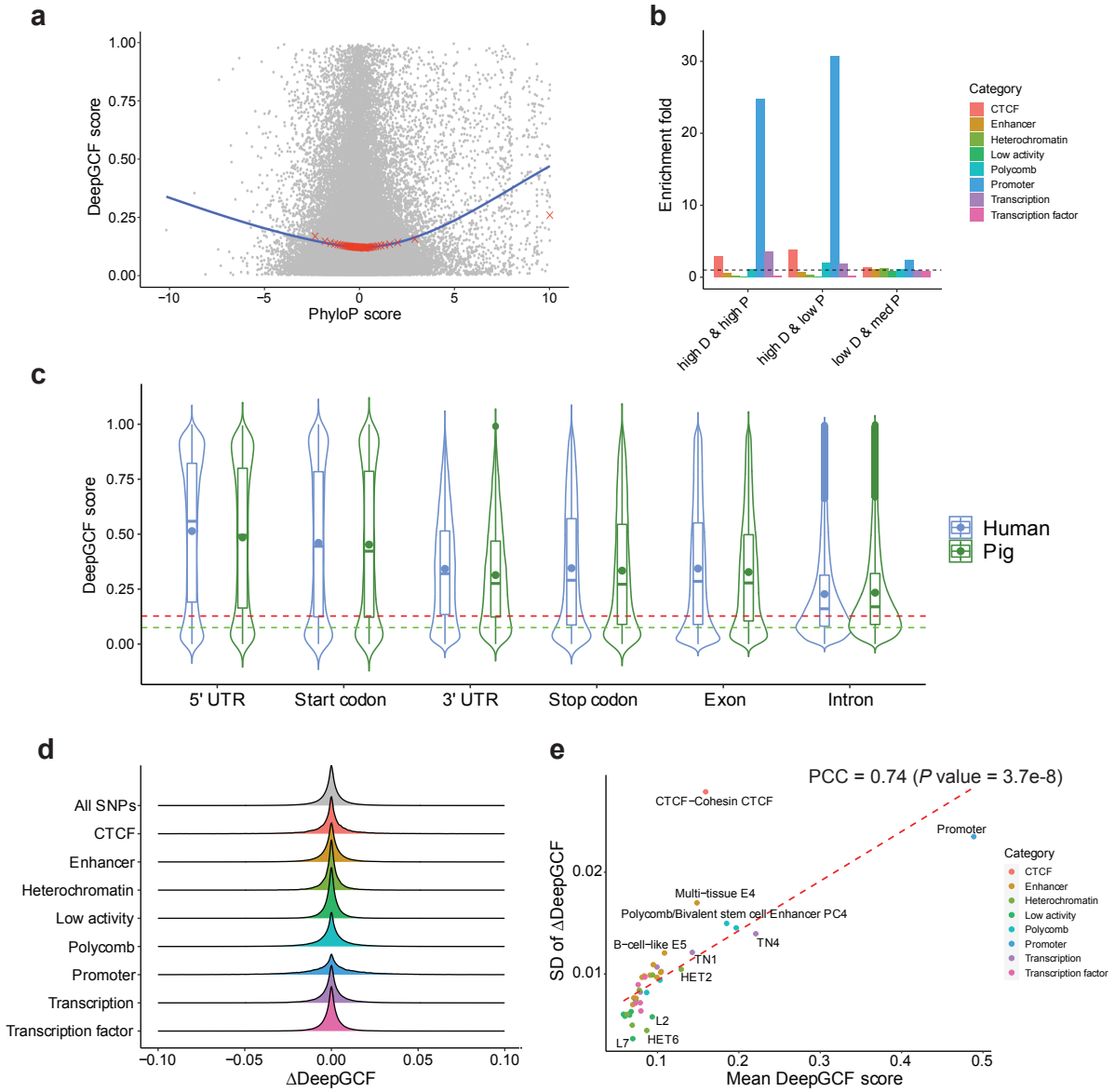

Fig. 3 Comparison of functional and sequence conservations. a Relationship between DeepGCF score and PhyloP score of 20,000 randomly selected human regions. PhyloP score is based on multiple alignments of 99 vertebrate genomes to the human genome[10]. The blue line is the fitted loess regression and red crosses represents 50 equally-divided percentiles of PhyloP score corresponding to the average of DeepGCF score. **b** Enrichment fold of 8 sequence class categories[28] for regions with high DeepGCF (> 95th percentile) and high PhyloP (> 95th percentile; high D & high P; n = 260,281), regions with high DeepGCF but low PhyloP (< 5th

632 percentile; high D & low P; n = 152,557), and regions with low DeepGCF (< 5th percentile) and
633 medium PhyloP (between 47.5th and 52.5th percentile; low D & med P; n = 77,848). Enrichment
634 is equal to the proportion of a sequence class category for a type of orthologous regions divided
635 by that for the whole genome. The dashed line (set at 1) represents no enrichment. **c** DeepGCF
636 score distribution of the different sequence ontologies. The red and green dashed lines represent
637 the mean and the median DeepGCF score of the whole genome. The dots inside each box
638 represent the mean DeepGCF score. **d** ΔDeepGCF (DeepGCF after mutation – original DeepGCF)
639 caused by 1,000,000 randomly selected orthologous variants, which are classified into 8 sequence class
640 categories[28]. The red dashed line represents the fitted regression line. **e** The effect of orthologous
641 variants (n = 35,575,835) on DeepGCF score of regions in 40 sequence classes[28], which are
642 classified into 8 categories. The effect was measured by ΔDeepGCF for variants in each
643 sequence class. The SD of ΔDeepGCF for each sequence class quantifies the overall sensitivity
644 of the sequence class to variant effect.
645

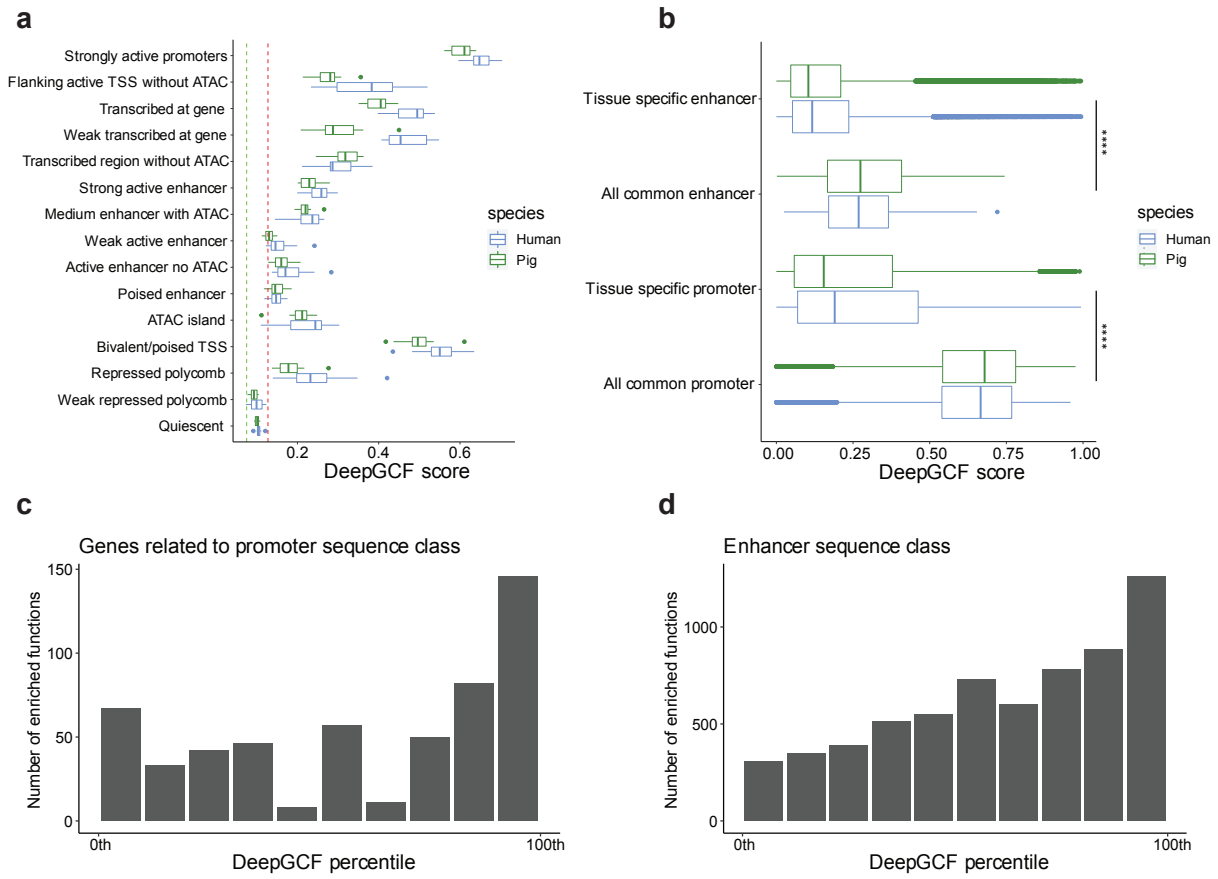

Fig. 4 DeepGCF score of genomic regions overlapping with different regulatory elements. a Distribution of average DeepGCF scores across human tissues (n = 12) and pig tissues (n = 14) for each chromatin state. The red and green dashed lines represent the mean and the median DeepGCF score of the whole genome. **b** DeepGCF scores of genomic regions overlapping with tissue-specific strongly active promoter and enhancer for human and pig[19]. "All common" represents promoters/enhancers shared across all tissues. **** denotes Mann–Whitney U test $P <$ 2.2e-16. **c** Number of significantly enriched gene ontology terms for human of genes related to promoters annotated by sequence class[28]. The genes were binned by DeepGCF into ten equal-width bins, and the functional enrichment analysis was conducted on each bin. **d** Similar to **c**, except showing the results of enhancers annotated by sequence class[28].

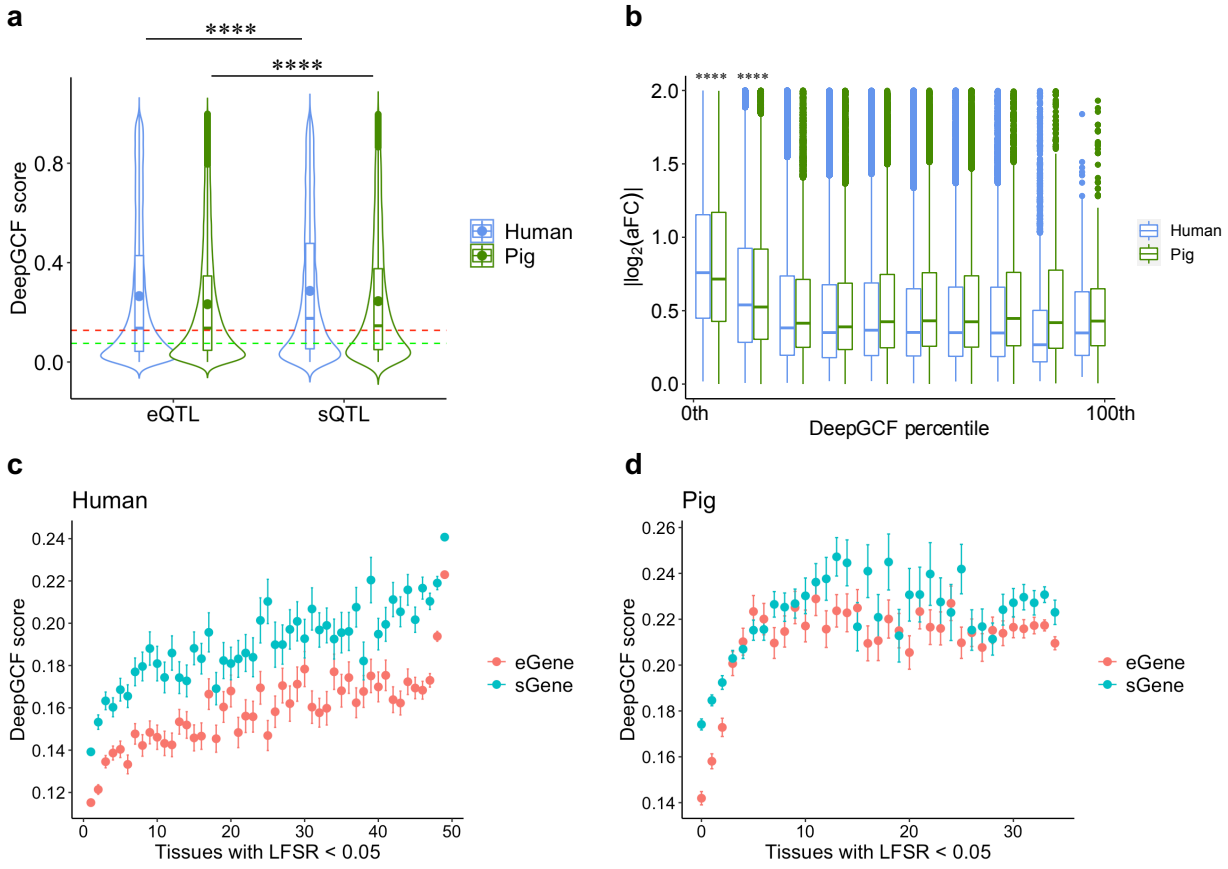

Fig. 5 Relationship of DeepGCF score to genetic variants. a The distribution of DeepGCF score of eQTLs and sQTLs. The red and green dashed lines represent the mean and the median DeepGCF score of the whole genome. The dots inside each box represent the mean DeepGCF score. **** denotes *P* value < 1e-8 based on a two-sided Mann–Whitney U test. **b** Relationship between the absolute value of eQTL effect size ($|\log_2(aFC)|$) and DeepGCF score for eGenes. The genes were binned by DeepGCF into ten equal-width bins for human and pig, respectively. **** denotes the group is different from all other groups with *P* value < 1e-8 based on a Tukey multiple comparison. **c** DeepGCF scores of tissue-sharing e/sGenes from human at local false sign rate (LFSR) < 5% obtained by MashR[50]. **d** Similar to **c**, except showing the results of pig.

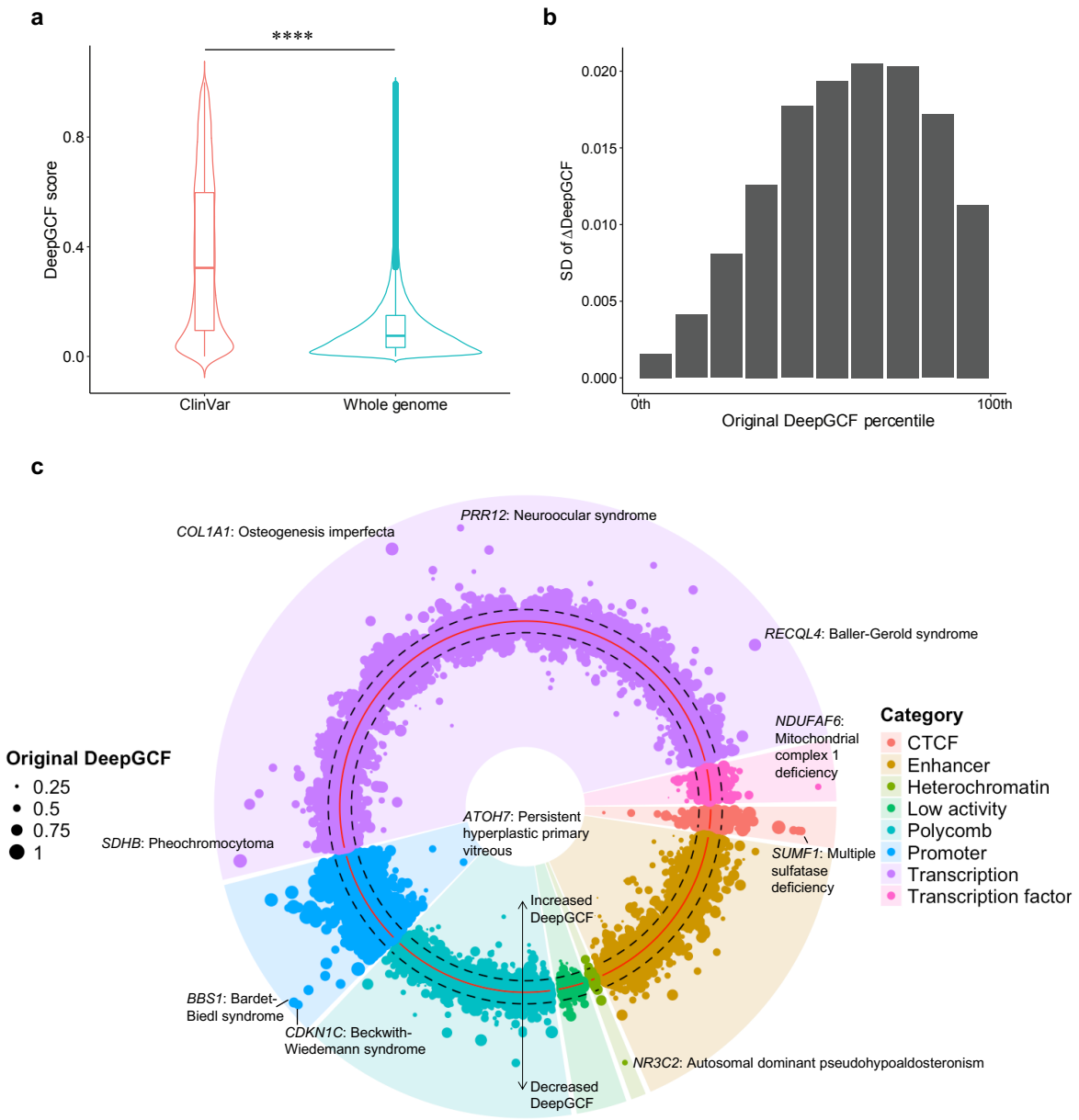

Fig. 6 Relationship of conservation score to pathogenic variants. a The distribution of DeepGCF scores in pathogenic and likely pathogenic SNPs (n = 104,033) obtained from ClinVar[38], compared to the DeepGCF distribution across the whole genome. **** denotes Mann–Whitney U test P < 2.2e-16. **b** SD of ΔDeepGCF (DeepGCF after mutation – original DeepGCF) caused by ClinVar SNPs. The SNPs were binned by their original DeepGCF into ten equal-width bins. **c** ClinVar SNPs classified by sequence class[28]. A polar coordinate system was used, where

675 the radial coordinate indicates the SNP effect on DeepGCF. The red solid circle represents zero
676 DeepGCF change, and two dashed circles represent ± 0.03 of DeepGCF encompassing 95% of
677 SNPs. Each dot represents a SNP and SNPs inside the red circle were predicted to have positive
678 effects (increased DeepGCF), while SNPs outside the red circle were predicted to have negative
679 effects (decreased DeepGCF). Dot size indicates the original DeepGCF. Within each sequence
680 class, SNPs were ordered by chromosomal coordinates. Top 10 SNPs with large impact on
681 DeepGCF associated disease and gene names were annotated.

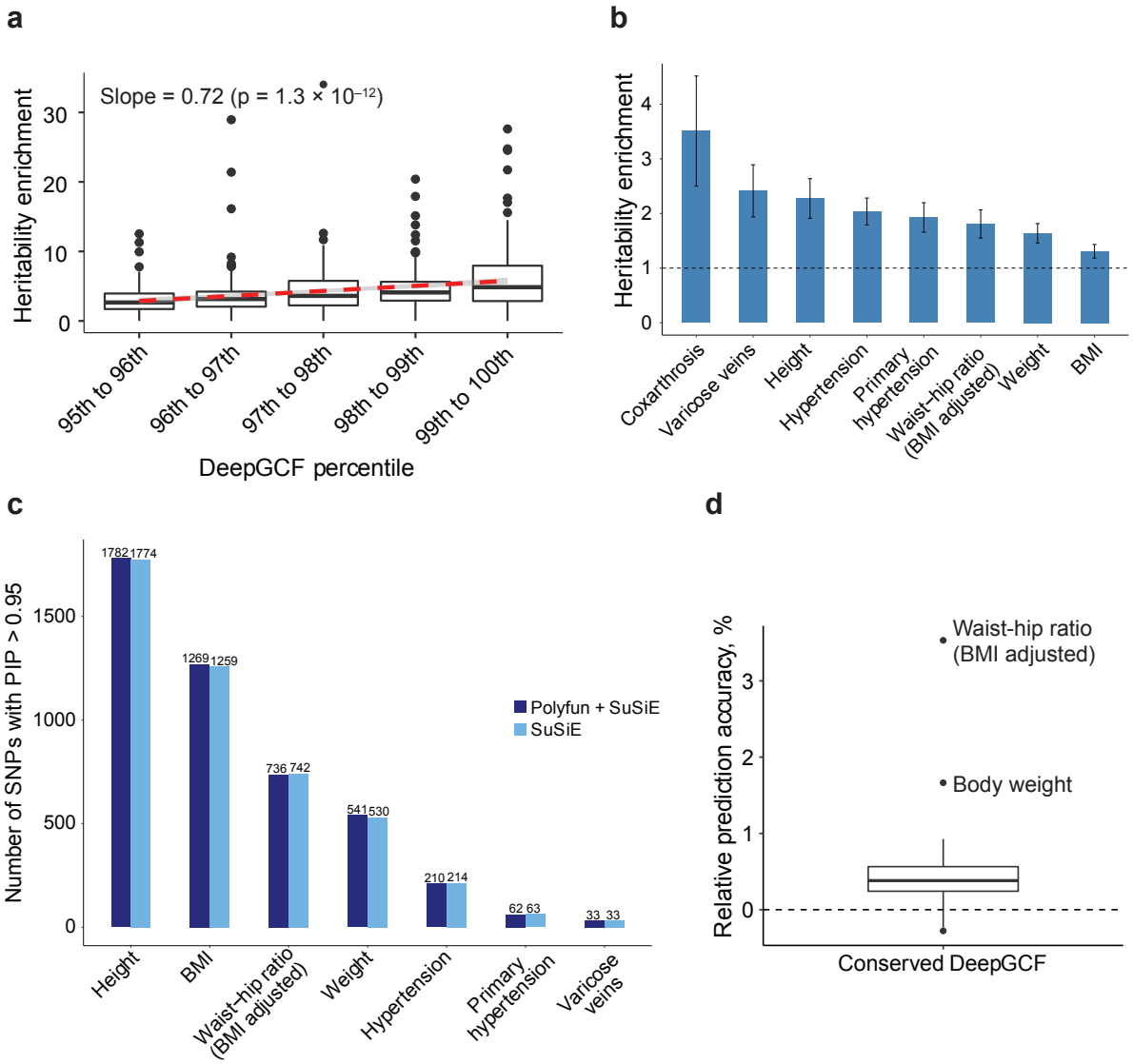

Fig. 7 Application of DeepGCF on complex traits/diseases in human. a Heritability enrichment calculated by LDSC for 80 human traits using functionally conserved regions (top 5% DeepGCF). The regions were divided into 5 equal equal-width bins and the heritability enrichment of all traits was calculated for each bin. The dashed red line is the fitted regression line between heritability enrichment and DeepGCF percentile, and the grey area is the 95% confidence interval. **b** Significant heritability enrichment explained by functionally conserved regions in 8 human traits. **c** The number of putatitive SNPs (PIP > 0.95 and P < 5e-8) identified by PolyFun + SuSiE using functionally conserved regions as a prior and SuSiE without priors for

691 7 human traits. **d** The relative prediction accuracy of PRS for 20 human complex traits using
692 functionally conserved regions as a prior in SBayesRC[43]. Relative prediction accuracy is equal to
693 (prediction accuracy using the prior – prediction accuracy without priors) / prediction accuracy
694 without priors. A relative prediction accuracy > 0 (dashed line) indicates an accuracy higher than
695 without priors.