**Supporting Information for**

**DeepAnnotation: A novel interpretable deep learning-based genomic selection model that integrates comprehensive functional annotations**

Wenlong Ma1,2, Shenghua Qin1,2, Weigang Zheng1,2,4, Chao Wang1,2, Bowen Lei1,2, and Yuwen Liu1,2,3,\*

1Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Key Laboratory of Livestock and Poultry Multi-Omics of MARA, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China.

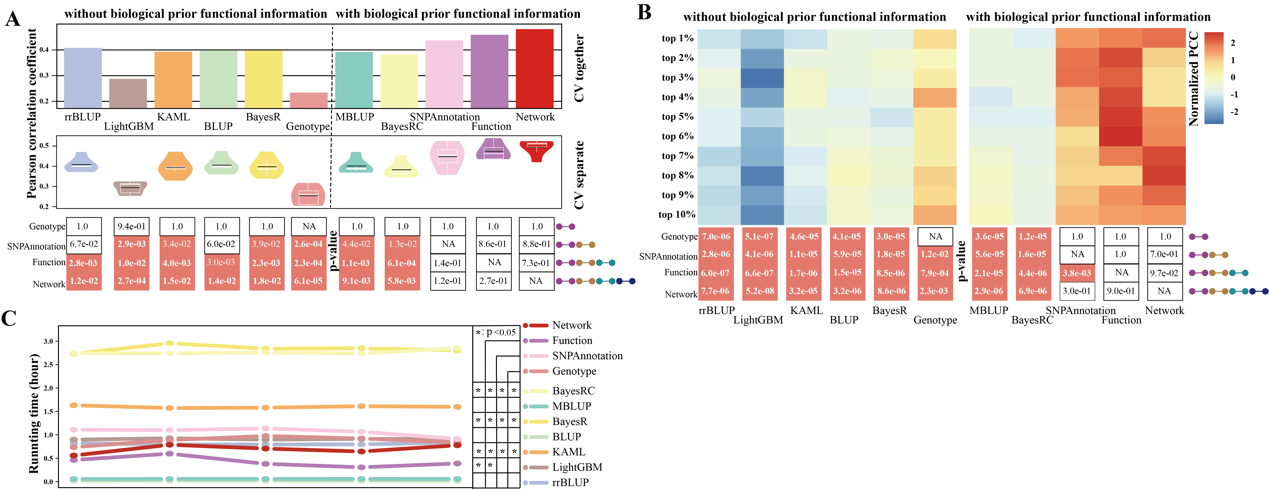
2Innovation Group of Pig Genome Design and Breeding, Research Centre for Animal Genome, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China.

3Kunpeng Institute of Modern Agriculture at Foshan, Chinese Academy of Agricultural Sciences, Foshan 528226, China.

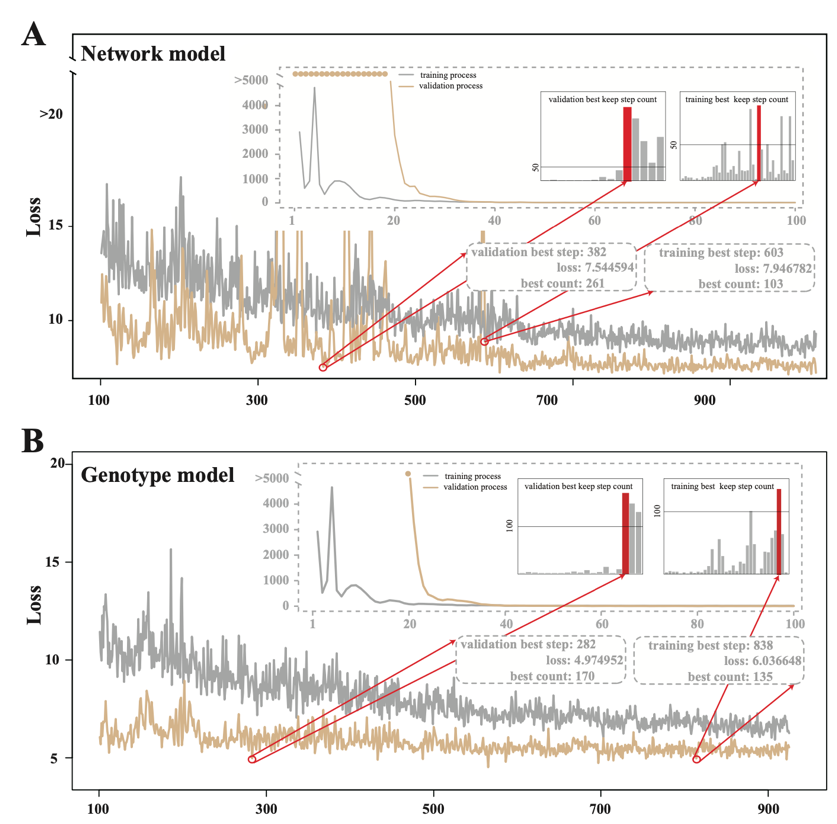
4Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education & Key Lab of Swine Genetics and Breeding of Ministry of Agriculture and Rural Affairs, Huazhong Agricultural University, Wuhan 430070, People’s Republic of China

\*Correspondence address: Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Key Laboratory of Livestock and Poultry Multi-Omics of MARA, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China; E-mail: liuyuwen@caas.cn

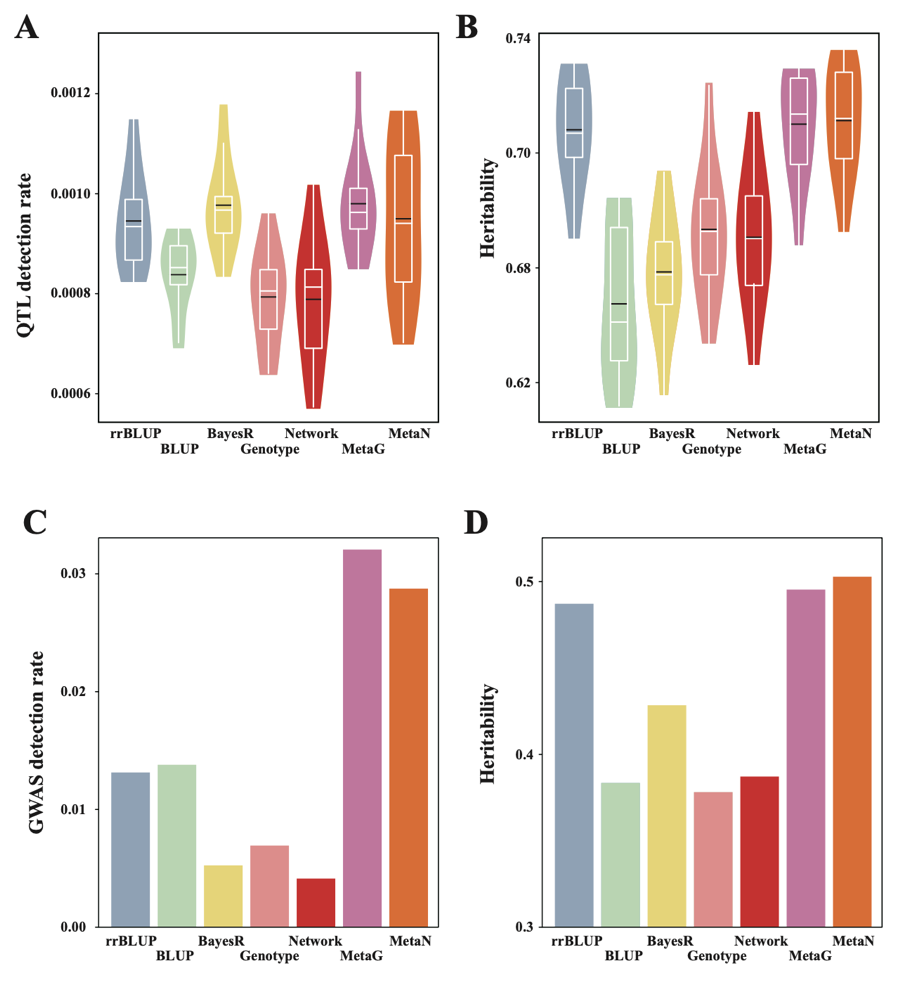
**Analyses**

**Dissecting the genetic basis of complex traits with DeepAnnotation**

Support Figure 1. Prediction performance of DeepAnnotation through cross-validation compared with rrBLUP, LightGBM, KAML, BLUP, BayesR, MultiBLUP, BayesRC, and DeepAnnotation on LMP trait. (A) PCC evaluation of different models based on 5-fold cross-validation experiment. In the "CV together" approach, each testing fold from the 5-fold cross-validation were consolidated before metric computation. On the other hand, in the "CV separate" approach, metrics were computed separately for each testing fraction of the 5-fold cross-validation. The paired t-test *P*-values of DeepAnnotation compared with other models were displayed on the black boxes with pink represents significance with *P*-value < 0.05. (B) Spatial PCC values between the predicted and observed phenotypic values of top-ranked samples from top 1% to top 30%. The paired t-test *P*-values were displayed on the black boxes with pink represents significance with *P*-value < 0.05. (C) Elapsed training time of different models.

The prediction performance evaluations showed that as the number of functional annotations incorporated into the DeepAnnotation models increased, the accuracy gradually improved (Support Figure 1). Considering that DeepAnnotation always performed better than other models in identifying individuals with top-ranked phenotype values, we were interested in whether DeepAnnotation could correctly dissect the causal variants which potentially dominate the genetic variance.

Support Figure 2. Training and validation status of Network (A) and Genotype (B) models of each epoch based on the pre-trained result from 5-fold cross-validation.

However, causal variants are usually hard to know. One way to do this is to simulate traits with pre-defined causal variants and check the detection rate. Therefore, we conducted this evaluation through 10 simulated traits with 10,000 pre-defined QTLs under 0.7 heritability (Methods). We selected Network and Genotype models to represent DeepAnnotation fully incorporated functional annotations or not, and trained Network and Genotype models at the total training epoch of 382 (Support Fig. 2A) and 282 (Support Fig. 2B). According to the performances and reasonable running time (Support Fig. 1B and 1D), we selected rrBLUP (adjusted BLUP-based model), original BLUP (no significant difference with MBLUP, paired t-test with *P*-value = 0.25) and BayesR (adjusted Bayesian-based model, no significant difference with BayesRC, paired t-test with *P*-value = 0.08) as baseline comparison models.

Support Figure 3. Dissecting the genetic basis of complex traits. (A) QTL detection rates distribution of different models on 10 simulated traits. (B) Estimated heritabilities distribution of significant SNPs (*P*-value < 0.05) identified by different models on 10 simulated traits. (C) GWAS significant SNPs (with adjusted *P*-value < 0.05) detection rates of different models on LMP. (D) Estimated heritabilities of significant SNPs (*P*-value < 0.05) identified by different models on LMP.

We found that the averaged QTL detection rates of the 10 simulated traits were 0.000793, 0.000838, 0.000946, and 0.000977 for Genotype, BLUP, rrBLUP, and BayesR, respectively (Support Fig. 3A). However, the averaged heritabilities explained by those significant SNPs (*P*-value < 0.05) driven by each model were 0.647, 0.658, 0.673, and 0.707 for BLUP, BayesR, Genotype, and rrBLUP, respectively (Support Fig. 3B). Although QTL detection rate of Genotype performed worst, the heritability of those significant SNPs explained by Genotype was better than BayesR (evaluated with the highest QTL detection rate), suggesting DeepAnnotation explained different aspects of genetic variance compared with other baseline models. Hence, we utilized meta-analysis strategy to combine all results together to test whether DeepAnnotation could serve as a complement to commonly used BLUP and Bayesian-based models in dissecting the genetic basic of complex traits. As expected, metaG (combination of rrBLUP, BLUP, BayesR, and Genotype) achieved the highest QTL detection rate of 0.000980 (Support Fig. 3A), and the highest estimated heritability of 0.709 (Support Fig. 3B). Furthermore, we calculated these two indices for Network to test whether inveracious functional annotations still helpful. The result showed that the overall QTL detection rate was 0.000790 and heritability was 0.671, which were both lower than Genotype (Support Fig. 3A-3B). Besides, metaN (combination of rrBLUP, BLUP, BayesR, and Network) was also displayed inconspicuous superiority than metaG (QTL detection rate: 0.000950 *vs* 0.000980, heritability: 0.711 *vs* 0.709), highlighting the importance of comprehensive and authentic functional annotations.

Another way to do this is to compare the predicted causal variants with GWAS results by assuming the SNPs with Bonferroni-adjusted *P*-value < 0.05 to be as potential causal variants. Therefore, we conducted this evaluation on real LMP trait (Methods). The same conclusion aline with the simulated traits could be reached under the indices of GWAS detection rate and heritability: the former were 0.00413, 0.00524, 0.00693, 0.0131, 0.0138, 0.0288, and 0.0321 for Network, BayesR, Genotype, rrBLUP, BLUP, metaN, and metaG, respectively (Support Fig. 3C); the latter were 0.378, 0.383, 0.387, 0.428, 0.487, 0.495, and 0.503 for Genotype, BLUP, Network, BayesR, rrBLUP, metaG, and metaN, respectively (Support Fig. 3D). Obviously, as a complement to commonly used models, DeepAnnotation actually improved the GWAS detection rate and estimated heritability.

In summary, benefiting from flexible framework of DeepAnnotation about incorporating comprehensive functional annotations and deep learning framework, DeepAnnotation may explained different aspects of genetic variance, and could be suggested as a completement to further increase the power of dissecting the genetic basis.

**Methods**

**Simulated data**

In this study, we used the simulated data to detect the genetic basis of complex traits by pre-defining causal variants. For each individual , the simulated phenotype was done by the following formula:

Where, represents the overall mean, represents the genotype vector of individual , represents the simulated effects vector of pre-defined QTLs, represents the residual effect. Here, represents the residual variance and the genetic variance , with represents the heritability In our simulations, 10 simulated traits were done based on the original 11,633,164 genotypes by setting the heritability to 0.7 and the number of pre-defined QTLs to 10,000.

**Detection rate and estimated heritability**

For each model, we firstly extracted 5 weights of all SNPs that were trained by each model through 5 CV experiment, and calculated their significant levels based on meta strategy with multiple testing correction through ‘RobustRankAggreg’ R package. The SNPs with adjusted *P*-value < 0.05 were consider as potential causal variants. For simulated traits, the real causal variants were considered as the pre-defined QTLs, while for real LMP trait, the real causal variants were considered as SNPs with Bonferroni-adjusted *P*-value < 0.05 from GWAS analysis (plink with sets ‘--linear –adjust’). Finally, the detection rate was defined as the ratio of real causal variants among the predicted causal variants.

The heritability was defined as the genetic variance explained by those predicted causal variants, and we estimated the heritability with BLUP model by setting ‘--reml-pred-rand --reml-est-fix --blup-snp’ with GCTA software.