GpNet: Genomic Prediction Network Using Locally Connected Layers in Korean Native Cattle

Hyo-Jun Lee
Chungnam National University
Dong Won Seo
Chungnam National University
Yoonji Chung
Chungnam National University
Doo Ho Lee
Chungnam National University
Yeung Kuk Kim
Chungnam National University
Jun Heon Lee
Chungnam National University
Hak Kyo Lee
ChonBuk National University
Cedric Gondro
Michigan State University
Young-Kuk Kim
Chungnam National University
Yeong Jun Koh
Chungnam National University
Seung Hwan Lee (slee46@cnu.ac.kr)
Chungnam National University

Research Article

Keywords: Genomic prediction, Deep learning, GWAS

Posted Date: June 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-622476/v1

RESEARCH

GpNet: Genomic prediction network using locally connected layers in Korean native cattle

Hyo-Jun Lee[1], Dong Won Seo[2], Yoonji Chung[2], Doo Ho Lee[2], Yeung Kuk Kim[2], Jun Heon Lee[2], Hak Kyo Lee[3], Cedric Gondro[4], Young-Kuk Kim[5], Yeong Jun Koh[5*] and Seung Hwan Lee[2*]

[*]Correspondence: yjkoh@cnu.ac.kr; slee46@cnu.ac.kr

[5] Department of Computer Science & Engineering, Chungnam National University, 305-764, Daejeon, Korea

[2] Division of Animal and Dairy Science, Chungnam National University, 305-764, Daejeon, Korea

Full list of author information is available at the end of the article

Abstract

Background: The use of DNA marker information for the prediction of genetic merit in animal and plant breeding, and susceptibility to disease in human medicine has become widespread. Therefore, an increasing number of methods have been proposed for more accurate and efficient genomic prediction. However, most of the commonly used models for genomic prediction only account for additive effects since most of them are designed based on the linear model.

Results: Here, we proposed a GpNet, a deep learning network for genomic prediction in Korean beef cattle. With a locally connected layer, GpNet can estimate LD-block effects of single nucleotide polymorphisms (SNP) with adjacent two or more SNPs closer to 3'-end. This operation is quite similar to how the DNA sequence is used in the translation process in which the RNA polymerase interprets DNA sequence by units of codons to downstream (3' to 5'). GpNet archived a superior performance than previous state-of-arts methods for beef carcass weight with a predictive ability of 0.721%. GpNet also found two significant quantitative trait locus (QTL) on the regions (bta 6:38464203–39816133, bta 14:25307116–29987025) for carcass weight. However, GpNet showed less performance than linear methods in backfat thickness and eye-muscle area.

Conclusions: GpNet outperformed the previous state-of-arts methods for beef carcass weight. However, GpNet cannot achieve superior performance in backfat thickness and eye-muscle area. We noticed that the lack of ability to estimate distant epistasis and dominance was the weakness of GpNet. Therefore, it remains a future research issue to expand GpNet to resolve these flaws and this further study will accelerate the new phase of the genomic prediction.

Keywords: Genomic prediction; Deep learning; GWAS

Background

The use of DNA marker information for the prediction of genetic merit in animal and plant breeding, and susceptibility to disease in human medicine has become widespread. This genomic information has been utilized primarily to detect regions of the genome that have an association with a specific phenotype (genome-wide association studies – GWAS) or to predict the genetic merit and phenotypes of individuals (genomic prediction) with many thousands of DNA markers, most commonly single nucleotide polymorphisms (SNP), covering the entire genome. In humans, genomic prediction has been widely used to predict disease risk and highly polygenic complex human traits [1, 2]. In agriculture, genomic prediction was used to estimate genomic breeding values (gEBV) which are then used to make selection decisions in a breeding population.

Most of the commonly used models for genomic prediction have been proposed based on the linear mixed models [3, 4]. Genomic best linear unbiased prediction (GBLUP) uses a mixed model approach which approximates a traditional infinitesimal model and assumes all SNP contribute a non-zero value to the genetic variance [4]. It is a method that simply uses a genomic relationship matrix built from the genotypes instead of a traditional pedigree-based relationship matrix. Bayesian linear model assumes that some SNPs have zero effects, whereas others have small to moderate effects and uses the posterior distributions to the parameters of linear mixed model [3, 5]. Even though these methods showed the state-of-the-art performance on many populations, they only account for additive effects, since most of them are designed based on the linear model. Thus, extended methods to account for non-linearity effects, such as dominance and epistatic interactions, have been proposed recently [6, 7].

Deep learning is also a good alternative method to solve this problem. Recent advances in deep neural networks have outperformed the state-of-the-art in computer vision, natural language processing, and audio recognition tasks [8, 9, 10, 11]. Using the local information of the input features, like image RGB-channel, text, or audio sequence, accelerated the successes of deep neural networks. Convolutional neural network (CNN), which is the most successful deep learning structure in computer vision, constitutes weights-shared filter operation for the adjacent region of input image [12]. Recurrent neural network (RNN) has been commonly used in sequence

to sequence problems, such as speech to text or natural language processing, generating a new sequence of the specific time by using the information before that time of sequence [10]. These two networks hypothesize that the regions showing similar patterns in the input data could explain the similar features with each other. As shown in Fig 1(a), features in the image (*e.g.* hair, eye, nose, glass, and so on) have similar RGB-color patterns within the same features. Speech sound also has a similar frequency pattern with other similar sounds (Fig 1(b)).

Figure 1 Example of image and sound data. (a) RGB-image; (b) Mel spectrogram of raw sound sequence.

Interestingly, the local information also can be addressed in the genomic prediction. The general concept of genomic prediction relies on the linkage disequilibrium (LD) between genetic markers and the unknown quantitative trait loci (QTL). With high-density SNP panels, the markers co-segregate with the causal mutations allowing their genetic effects to be indirectly estimated through the adjacent SNPs [3, 13]. Considering this attribute of SNP data, genomic prediction model should estimate the effects of each LD-block consisting of locally adjacent two or more SNPs not a single SNP for the more accurate prediction. However, unlike the image and sound data, the LD-blocks even with the same SNP pattern do not always have the same effect on the individual traits. For the SNP data, it is more important to recognize how close each LD-block biologically to the unknown QTL than the SNP pattern. Therefore, a different approach from the previous deep learning networks, such as CNN or RNN, is required to use local information for genomic prediction. Practically, the simple fully-connected networks that didn't use the local information usually showed better performance than other local-based networks in previous studies [14]. In addition, Zingaretti *et al.* [15] explored a convolutional neural network for genomic prediction of polyploid outcrossing species. Montesinos-López *et al.* [16] used deep learning to the multi-environment genomic prediction of plat complex traits. Pook *et al.* [17] applied the local convolutional neural networks on simulated maize and real Arabidopsis data. However, these studies can not provide clear evidences that deep neural networks can outperform the previous methods such as GBLUP or Bayesian linear models.

In this study, we proposed the Genomic prediction Network (GpNet) using a locally connected layer for genomic prediction in Korean native cattle. The locally connected layer works similarly to the causal convolution, except that weights are unshared, that is, a different set of weights is applied at each different LD-block. We validated the performance of GpNet as follow processes. First, the GpNet performances were evaluated on carcass weights, backfat thickness, and eye-muscle area of Korean native cattle, and then its performance was compared with the GBLUP [4], BayesA [3], and BayesLASSO [18]. Second, we also identified the candidate QTL region using LD-block effects estimated by GpNet for each trait. Since there are few results that deep learning outperformed the linear method, this study will be a very interesting attempt in the field of genomic prediction.

Results

Model performance

Table 1 presents the performance of GpNet, GBLUP, BayesA and BayesLASSO. In Table 1, we saw that GpNet (0.721) outperformed other linear method (GBLUP: 0.714, BayesA: 0.719, BayesLASSO: 0.712) in carcass weight (CWT). However, GpNet showed less performance than the linear methods in backfat thickness (BF) and eye-muscle area (EMA). Comparing between linear methods, BayesA showed the best performance in BF (0.637) and EMA (0.728).

Table 1 Predictive ability of each model.

	CWT	BF	EMA
GpNet	**0.721 ± 0.018**	0.602 ± 0.01	0.708 ± 0.011
GBLUP	0.714 ± 0.018	0.626 ± 0.011	0.727 ± 0.014
BayesA	0.719 ± 0.019	**0.637 ± 0.012**	**0.728 ± 0.013**
BayesLASSO	0.712 ± 0.02	0.629 ± 0.011	0.723 ± 0.014

An LD-pruned SNP set was also used in this study. Briefly, pairs of SNPs in the 1000-kb with a squared correlation greater than 0.1 were noted, and these SNPs were greedily pruned from the window until no such pairs remained. Finally, a total of 21,629 SNPs was used as an LD-pruned SNP set. Table 2 shows the predictive ability of each model with LD-pruned SNP. With the results using 50k SNP, GpNet once again showed the best performance (0.712) in the CWT. However, in BF and EMA, GpNet once again underperformed (BF: 0.589, EMA: 0.704) than linear methods, and GBLUP showed the best performance (BF: 0.607, EMA: 0.72) for

these traits. Comparing with the 50K SNP results, the predictive abilities of all models were decreased. These results corroborate the ideas of Manolio *et al.* [19], who maintained that a marker subset may cause the missing heritability even though the variants in subset can explain a large proportion of genetic variance.

Table 2 Predictive ability of each model with LD-pruned SNP.

	CWT	BF	EMA
GpNet	**0.712 ± 0.022**	0.589 ± 0.015	0.704 ± 0.013
GBLUP	0.701 ± 0.019	**0.607 ± 0.009**	**0.72 ± 0.016**
BayesA	0.709 ± 0.019	0.605 ± 0.01	0.718 ± 0.015
BayesLASSO	0.704 ± 0.019	0.606 ± 0.008	0.718 ± 0.015

Identification of QTL using GpNet

To compare the QTL mapping of GpNet, we also estimated the SNPs effect using the single marker linear mixed model (SMLMM). Fig 2 shows the mapping results of two methods (GpNet and SMLMM). In the CWT (Fig 2(a)), two significant peaks (bta 6:38464203–39816133 and bta 14:25307116-29987025) were found in both methods and one peak (bta 4:4508164–4790444) was identified only by SMLMM. All three regions were previously identified as QTLs for carcass weights in Korean beef cattle [20, 21, 22, 23]. As shown in Fig 2(b), the genetic characteristic of BF seemed to be more polygenic than CWT. Among the numerous significant peak, we saw that the five loci (bta 2:99845066-107657675, bta 6:38464203–39816133, bta 13:53003864-54829615, bta 19:6241437-7833508, and bta 23:3320932-4781751) were standing out at both methods. GpNet found a variant on bta 10:12412780 as the most significant marker for BF. This variant is on the protein-coding region of *DPP8* gene that was reported to attribute the dipeptidyl peptidase activity of cow testis [24]. EMA seemed to have a similar genetic structure with CWT (Fig 2(c)). These results seemed to be due to the genetic correlation between EMA and CWT. In our data, EMA showed the 0.546 correlation with CWT. We can see that variant on bta 27:23040097, which was not identified at CWT, was significant to EMA. This variant is close to *DLC1*, which plays a key role in the regulation of small GTP-binding proteins.

Table 3 shows the QTL region identified by both GpNet and SMLMM. In the results, *SLIT2* seemed to be a key gene for complex traits of Korean beef cattle since the association of *SLIT2* was replicated for CWT, BF, and EMA. *SLIT2* play

Figure 2 Manhattan plots of SMLMM and GpNet for each trait. (a) carcass weight; (b) backfat thickness; (c) eye-muscle area; x-axis is SNP position and chromosome, y-axis is $-\log_{10}$ *P-value*.

highly conserved roles in axon guidance and neuronal migration. A lot of genome-wide studies have reported this gene to QTL for beef complex traits including organ weight [25, 26], body weight [27], and fertility [28].

Table 3 Predictive ability of each model with LD-pruned SNP.

Trait	Region	Close gene
CWT	4:4508164-4790444	*COBL*
	6:38464203-39816133	***SLIT2***
	14:25307116-29987025	*RAB2A*
BF	2:99845066-107657675	*SLC4A3*
	6:38464203–39816133	***SLIT2***
	13:53003864-54829615	*NTSR1*
	19:6241437-7833508	*C19H17orf67*
	23:3320932-4781751	*BMP5*
EMA	6:38464203–39816133	***SLIT2***
	10:20359799-21698931	*DHRS4*
	11:68687376-69190084	*LCLAT1*
	12:32687103-33576827	*LOC536660*
	14:25307116-29987025	*RAB2A*

Discussion

The flexibility of a deep neural network takes advantage of its non-linearity for genomic prediction in comparison to the traditional linear-based methods. Our proposed GpNet can explore the epistasis of adjacent SNPs, called the LD-block effect in this study, using a locally connected layer. This structure may stand to benefit for the trait with some obvious causal loci, like CWT in this study. However, GpNet under-performs than linear methods for BF and EMA, which have more polygenic structure than CWT. In the polygenic trait, a lot of loci at various distances make epistasis, which is quite challenging since GpNet only considers the interaction between adjacent SNPs. As shown in Fig 2, QTL mapping of GpNet for BF and EMA was unclear, compared to SMLMM.

For capturing distant epistasis, a dilated convolution [29] would be a good alternative. It is equivalent to a convolution operation with a larger kernel filter by dilating it with zeros. In the WaveNet [9], Stacked dilated convolution enabled networks to have very large receptive fields for raw audio sequence. Otherwise, a dilated convo-

lution can estimate very large epistasis fields for genomic prediction. Even though a locally connected layer included a dilated operation, it only consisted of a very low delation rate due to the memory complexity $O(n(d+1))$, where n is the number of SNPs and d is depths of locally connected layer (Fig 3). On the other hand, dilated convolution effectively allows the networks to estimate epistasis with a much large distance since it only requires the two shared parameters per layer (Fig 4). Therefore, it remains a future work to incorporate dilated convolution to GpNet for more accurate genomic prediction. Specifically, a locally connected layer at GpNet can be used to estimate adjacent epistasis, while additional modules with dilated convolution estimate distant epistasis. Then, the final epistasis can be accounted by a combination of these two estimated values.

Figure 3 Visualization of the locally connected layer. n is the number of input SNPs, d means layer depths. Since weights are unshared at locally connected layer, the number of parameters at d-depth layer is $(d+1)n$

Figure 4 Visualization of dilated convolution. n is the number of input SNPs, d means layer depths. Since weights are shared at dilated convolution, the number of parameters at each layer is two.

Dominance also contributes to the total genetic potential for the phenotype (Fig 5(a)). A nonlinear activation function, which is a critical part of the design of a neural network, can allow such networks to compute nontrivial dominance. GpNet adopts relu as a nonlinear activation function. However, relu is not suitable for identifying dominance since it is still linear for positive values. Instead of relu, transformed sigmoid or tanh would be good options for calculating dominance (Fig 5(b)).

Figure 5 Type of dominance and candidate activations for dominance. (a) Type of dominance; (b) Candidate activations for dominance.

Our proposed GpNet in this work did not consider the distant epistasis and dominance. Fig 6 shows the identical model for genomic prediction. Even though some candidate methods to implement each module (additive, epistasis, and dominance) were discussed in this study, designing a full model in Fig 6 is of course still chal-

lenging. The optimal architectures (depth and width) of each module, which is dependent on the other modules' architectures, should be found separately and the aggregate methods to combine the features transformed from each module must be determined. In addition, the interaction of the three modules is also needed to be implemented in the full model. Therefore, it remains a future research issue to expand GpNet to an ideal genomic prediction model in Fig 6. This further research will accelerate the new phase for genomic prediction.

Figure 6 Identical model and candidate methods for different type of genetic effects.

Conclusions

In this paper, we presented a GpNet, a deep learning network for genomic prediction in Korean native cattle. With a locally connected layer, GpNet can estimate genetic effects of each LD-block consisting of neighboring two or more SNPs. In the results, the GpNet outperformed previous state-of-arts methods including GBLUP, BayesA, BayesLASSO in CWT. However, GpNet can't achieve superior performance than linear models in BF and EMA. Furthermore, GpNet did not consider the distant epistasis and dominance effects. To resolve these flaws, we discussed alternative methods in **Discussion** section. With these alternatives, it remains a future research issue to expand GpNet to an ideal genomic prediction model and this further research will accelerate the new phase for genomic prediction.

Materials and methods

Dataset

The commercial Korean native cattle population used in this study included 10000 individuals (animals were born between 2010 $\sim$ 2017 and samples were collected between 2013 $\sim$ 2019) with phenotypic measurements for carcass weight (CWT/kg), eye-muscle area (EMA/cm^2), and backfat thickness (BF/mm). BF and EMA were measured after a 24-hour chill at the junction between the 12th and 13th ribs.

Genomic DNA of the animals was extracted from longissimus thoracis muscle samples using a DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA). A total of 10000 samples were genotyped using the Illumina Bovine SNP50 BeadChip. SNP quality control was performed using PLINK1.9 software [18] based on the following

filtering criteria: minor allele frequency < 0.001; SNP call rate < 0.1; SNP on the sex chromosomes. 1,853 SNP were excluded by the quality control filtering step and the postfilter missing rate was 0.6% of the genotypes. These missing SNP were then imputed with Eagle v2.4 [30] and a final total of 44,314 SNP were used in the study. The dataset was split into train (80%), validation (10%), and test (10%) sets to evaluate GpNet performance.

Notice that the National Institute of Animal Science (NIAS) in Rural Development Administration (RDA) of South Korea approved the experimental procedures, and all samples were taken under public animal health and welfare guidelines.

GpNet

In this paper, we proposed a new genomic prediction model operating on the SNP data. We can write the SNP data set as a one-dimensional sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ by base pair position. The goal of our proposed networks was to assign an LD-block effect (LDB) to SNP on the i-position (x_i) as follows:

$$LDB_{x_i} = \sum_{j=i-k}^{i} w_j^{(i)} x_j \tag{1}$$

where, x_j is genotype of SNP on the j-position; $w_j^{(i)}$ is LD-effect of SNP x_j to SNP x_i; k is the LD-window size. Therefore, the assigned LD-block effect LDB_{x_i} is conditional estimated with a total of k-SNPs closer to 3'-end. This operation is quite similar to how DNA sequence is used in the translation process in which the RNA polymerase interprets DNA sequence by units of codons to downstream (3' to 5') and generates mRNA sequences with this information. To model this operation, we opted a locally connected layer.

Locally connected layer

Fig 3 shows the visualization of the locally connected layer. A locally connected layer was inspired by causal convolution [9] and local convolution [31]. By using a locally connected layer, the network cannot violate the order of SNP. Otherwise, The LD-block effect of SNP at the i-position cannot depend on any of the SNPs to the 5'-end $(x_{i+1}, x_{i+2}, \ldots, x_n)$. A locally connected layer adopts a dilated operation where the window size is increased over layer depth. By this operation, a model can

estimate the LD effects from a wider LD-block as the depth of the networks get deeper.

Netwrok Structure

GpNet consists of the stacks of multiple locally connected layers (Fig 7). Both skip connection [11] and relu activation [32] are used throughout the network to enable training of a much deeper model. Given this structure, the model will estimate new LD-block effects at each layer and then add them to the input SNPs. We scaled the different layer depths d and stack number s for three different traits with a validation set. In particular, we found the best values for CWT (d:4, s:1), BF (d:4, s:3) and EMA (d:3, s:7). Finally, from the last locally connected layer, one fully-connected layer with non-activation (linear operation) wields a scalar, which is correspond to the genomic estimated breeding value (gEBV).

Figure 7 GpNet architecture. LCL is locally connected layer; d and s is the layer depths and the number of stacks; *Relu* is relu activation; GpNet can be scaled with d and s for different traits.

Loss function

In the common deep learning approach, the model aims to predict a truth label y from input sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$. However, the main purpose of genomic prediction is to estimate a genetic portion of individual phenotype, not truth phenotype itself. Therefore, we added the h^2 term in the mean squared error loss. Let σ_G^2 and σ_P^2 are genetic variance and phenotype variance. Then, the heritability can be calculated by $h^2 = \sigma_G^2/\sigma_P^2$. Finally, the loss function for training GpNet was defined as follows:

$$\mathcal{L}\left(y, \widetilde{y}\right) = \frac{\sum_{p=1}^{m} \left(h^2 y_p - \widetilde{y}_p\right)^2}{m} \tag{2}$$

where, y_p and $\widetilde{y}_p$ are observed phenotype and predicted phenotype of p-animal. By shrinking y with h^2, $\widetilde{y}$ will be converged to the gEBV at the end of model training. The variance component, σ_G^2 and σ_P^2, were estimated using an average information restricted maximum likelihood [33] by implementing the AIREMLF90 program [34]. Table 4 shows the variance estimation results for each trait.

Table 4 Variance component estimation results.

	CWT	BF	EMA
σ^2_G	962.3	9.5	52
σ^2_P	1495.6	15.5	90.2
h^2	0.392	0.378	0.366

Implementation details

For training GpNet, we set a learning rate to 10^{-4}. The training was iterated for 100 epochs with batch size 16 on GeForce RTX 3090. We employed Adam [35] optimizer to minimize the loss function. We determined network parameters, which achieved the best performance on the validation set among all epochs. To validate the GpNet ability, performances of GBLUP, BayesA and BayesLASSO were compared with GpNet results. All test steps were repeated 5 times, with randomly split train (80%), validation (10%), and test (10%) set. Since the linear methods don't need the validation set, the 9000 animals (train set + validation set) were used for training each linear model. We measured the model performance as the correlation between phenotype and gEBV divided by the square root of heritability, $cor(y, gEBV)/h$, called predictive ability [36].

Finding QTL with LD-block effects.

In addition to predict gEBV, GpNet also can be used for estimating the LD-block effect of each SNP. As the layer feedforwards to the next layer, GpNet accumulates the LD-effect of the SNP on the i-position (x_i) as follows:

$$f_d\left(x_i\right) = \sum_{j=i-k}^{i} w_j^{d(i)} f_{d-1}(x_j) + f_{d-1}(x_i) \tag{3}$$

where, f_d means the operation of the d-th layer; $w_j^{d(i)}$ is LD-effect of SNP x_j to SNP x_i on the d-th layer. Therefore, the final LD-block effects of SNP x_i can be defined as:

$$LD_{x_i} = f_E(x_i) - f_{E-1}(x_i) = \sum_{j=i-k}^{i} w_j^{E(i)} f_{E-1}(x_j) \tag{4}$$

where, f_E means the operation of the last locally connected layer; LD_{x_i} is final LD-block effects of SNP x_i. Then, LD-block effect can be calculated by the difference between the outputs of the last layer $f_E(x_i)$ and the one before it $f_{E-1}(x_i)$.

This LD-block effect, differing with the single SNP-effect estimated from the whole population, can be different values for each individual since the LD-block effects are estimated from individuals' SNPs patterns. To estimate the LD-block effects in this study, we trained the whole population (10000 animals) to GpNet, and then 1000 animals with high-gEBV and low-gEBV were noted (Fig 8). We hypothesized that the difference in gEBV ranking between these two groups (high and low) would be reflected by the difference in LD-block effects of each individual. Therefore, we did a t-test for finding a significant region as follows:

$$Sig_i = \text{t-test}\left(H_{x_i}, L_{x_i}\right) \tag{5}$$

where Sig_i is significant value of i-position; $H_{x_i} \in \mathbb{R}^{1000}$ and $L_{x_i} \in \mathbb{R}^{1000}$ are LD-block effects of SNP x_i in the high-gEBV group and low-gEBV group, separately.

Figure 8 The process of QTL mapping.

Abbreviations

- **SNP:** Single nucleotide polymorphisms
- **gEBV:** Genomic breeding values
- **GBLUP:** Genomic best linear unbiased prediction
- **CNN:** Convolutional neural network
- **RNN:** Recurrent neural network
- **LD:** Linkage disequilibrium
- **QTL:** Quantitative trait locus
- **CWT:** Carcass weight
- **BF:** Backfat thickness
- **EMA:** Eye-muscle area
- **SMLMM:** Single marker linear mixed model

Declarations

Ethics approval and consent to participate

Notice that the National Institute of Animal Science (NIAS) in Rural Development Administration (RDA) of South Korea approved the experimental procedures, and all samples were taken under public animal health and welfare guidelines.

Availability of data and materials

All source code for this study is freely available for download at https://github.com/gywns6287/GpNet. Request for Genotype data can be made to Korea National Institute of Animal Science, Animal Genome & Bioinformatics Division (http://www.nias.go.kr/english/sub/boardHtml.do?boardId=depintro).

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not Applicable.

Funding

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2019R1F1A1057605). CG was also supported by the National Institute of Food and Agriculture (AFRI Project No. 2019-67015-29323).

Authors' contributions

Conceptualization, HJL and YJK; Data curation, HKL and JHL; Methodology, DWS, YJC; Formal analysis, HJL and DHL; sofware, HJL, YKK, YKK; Writing – original draft, HJL; Writing – review & editing, SHL, CG. All authors have read and approved the final manuscript.

Acknowledgements

All data (phenotypes and genotypes) used in this study were provided by the Next Generation Biogreen21 project (PJ01316903), RDA. This study is part of an initiative to develop new genomic selection models using Artificial Intelligence and Machine Learning for Hanwoo cattle. We acknowledge the Next Generation Biogreen 21 Program, RDA.

Author details

[1] Department of Bio-AI Convergence, Chungnam National University, 305-764, Daejeon, Korea. [2] Division of Animal and Dairy Science, Chungnam National University, 305-764, Daejeon, Korea. [3] Department of Animal Biotechnology, ChonBuk National University, Jeonju, Korea. [4] Department of Animal Science, Michigan State University, East Lansing, MI, USA. [5] Department of Computer Science & Engineering, Chungnam National University, 305-764, Daejeon, Korea.

References

1. Abraham G, Inouye M. Genomic risk prediction of complex human disease and its clinical application [Journal Article]. Current opinion in genetics & development. 2015;33:10–16.
2. de Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor [Journal Article]. PLoS Genet. 2013;9(7):e1003608.
3. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps [Journal Article]. Genetics. 2001;157(4):1819–1829.
4. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix [Journal Article]. Genetics research. 2009;91(1):47–60.
5. Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, Reich C, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels [Journal Article]. Journal of dairy science. 2012;95(7):4114–4129.
6. Martini JW, Gao N, Cardoso DF, Wimmer V, Erbe M, Cantet RJ, et al. Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE) [Journal Article]. BMC bioinformatics. 2017;18(1):1–16.
7. Da Y, Wang C, Wang S, Hu G. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers [Journal Article]. PloS one. 2014;9(1):e87666.

8. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR; 2019. p. 6105–6114.
9. Oord Avd, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. Wavenet: A generative model for raw audio [Journal Article]. arXiv preprint arXiv:160903499. 2016;.
10. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [Journal Article]. arXiv preprint arXiv:14061078. 2014;.
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.
12. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks [Journal Article]. Advances in neural information processing systems. 2012;25:1097–1105.
13. Kizilkaya K, Fernando R, Garrick D. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes [Journal Article]. Journal of animal science. 2010;88(2):544–551.
14. Bellot P, de Los Campos G, Pérez-Enciso M. Can deep learning improve genomic prediction of complex human traits? [Journal Article]. Genetics. 2018;210(3):809–819.
15. Zingaretti LM, Gezan SA, Ferrão LFV, Osorio LF, Monfort A, Muñoz PR, et al. Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. Frontiers in plant science. 2020;11:25.
16. Montesinos-López A, Montesinos-López OA, Gianola D, Crossa J, Hernández-Suárez CM. Multi-environment genomic prediction of plant traits using deep learners with dense architecture. G3: Genes, Genomes, Genetics. 2018;8(12):3813–3828.
17. Pook T, Freudenthal J, Korte A, Simianer H. Using local convolutional neural networks for genomic prediction [Journal Article]. Frontiers in genetics. 2020;11.
18. Park T, Casella G. The bayesian lasso [Journal Article]. Journal of the American Statistical Association. 2008;103(482):681–686.
19. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–753.
20. Li Y, Lee JH, Lee YM, Kim JJ. Application of linkage disequilibrium mapping methods to detect QTL for carcass quality on chromosome 6 using a high density SNP map in Hanwoo [Journal Article]. Asian-Australasian Journal of Animal Sciences. 2011;24(4):457–462.
21. Lee SH, Choi BH, Lim D, Gondro C, Cho YM, Dang CG, et al. Genome-wide association study identifies major loci for carcass weight on BTA14 in Hanwoo (Korean cattle) [Journal Article]. PLoS One. 2013;8(10):e74677.
22. Sudrajad P, Sharma A, Dang CG, Kim JJ, Kim KS, Lee JH, et al. Validation of single nucleotide polymorphisms associated with carcass traits in a commercial Hanwoo population. Asian-Australasian journal of animal sciences. 2016;29(11):1541.
23. Lee HJ, Chung YJ, Jang S, Seo DW, Lee HK, Yoon D, et al. Genome-wide identification of major genes and genomic prediction using high-density and text-mined gene-based SNP panels in Hanwoo (Korean cattle). PloS one. 2020;15(12):e0241848.
24. Dubois V, Ginneken CV, Cock HD, Lambeir AM, Veken PVd, Augustyns K, et al. Enzyme activity and immunohistochemical localization of dipeptidyl peptidase 8 and 9 in male reproductive tissues [Journal Article]. Journal of Histochemistry & Cytochemistry. 2009;57(6):531–541.
25. An B, Xia J, Chang T, Wang X, Miao J, Xu L, et al. Genome-wide association study identifies loci and candidate genes for internal organ weights in Simmental beef cattle. Physiological genomics. 2018;50(7):523–531.
26. Raza SHA, Khan S, Amjadi M, Abdelnour SA, Ohran H, Alanazi KM, et al. Genome-wide association studies reveal novel loci associated with carcass and body measures in beef cattle. Archives of Biochemistry and Biophysics. 2020;p. 108543.
27. Smith JL, Wilson ML, Nilson SM, Rowan TN, Oldeschulte DL, Schnabel RD, et al. Genome-wide association and genotype by environment interactions for growth traits in US Gelbvieh cattle. BMC genomics. 2019;20(1):1–13.
28. Höglund JK, Buitenhuis B, Guldbrandtsen B, Lund MS, Sahana G. Genome-wide association study for female

fertility in Nordic Red cattle. BMC genetics. 2015;16(1):1–11.

29. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence. 2017;40(4):834–848.
30. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel [Journal Article]. Nature genetics. 2016;48(11):1443.
31. Taigman Y, Yang M, Ranzato M, Wolf L. Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014. p. 1701–1708.
32. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Icml; 2010. .
33. Meyer K. An'average information'restricted maximum likelihood algorithm for estimating reduced rank genetic covariance matrices or covariance functions for animal models with equal design matrices [Journal Article]. Genetics Selection Evolution. 1997;29(2):1–20.
34. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee D, et al. BLUPF90 and related programs (BGF90). In: Proceedings of the 7th world congress on genetics applied to livestock production. vol. 33; 2002. p. 743–744.
35. Kingma DP, Ba JL. Adam: A method for stochastic gradient descent. In: ICLR: International Conference on Learning Representations; 2015. p. 1–15.
36. Legarra A, Robert-Granié C, Manfredi E, Elsen JM. Performance of genomic selection in mice [Journal Article]. Genetics. 2008;180(1):611–618.

Figures

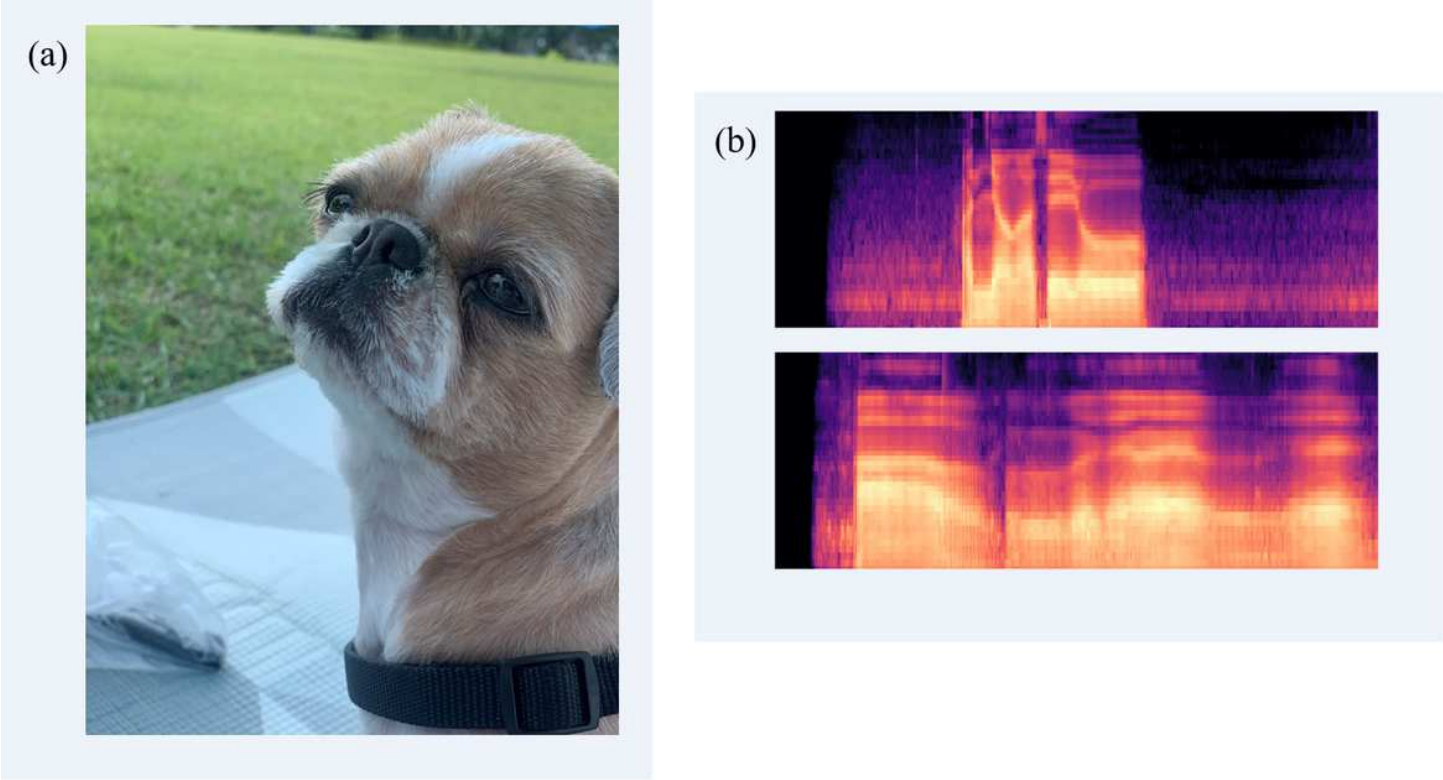

Figure 1

Example of image and sound data. (a) RGB-image; (b) Mel spectrogram of raw sound sequence.

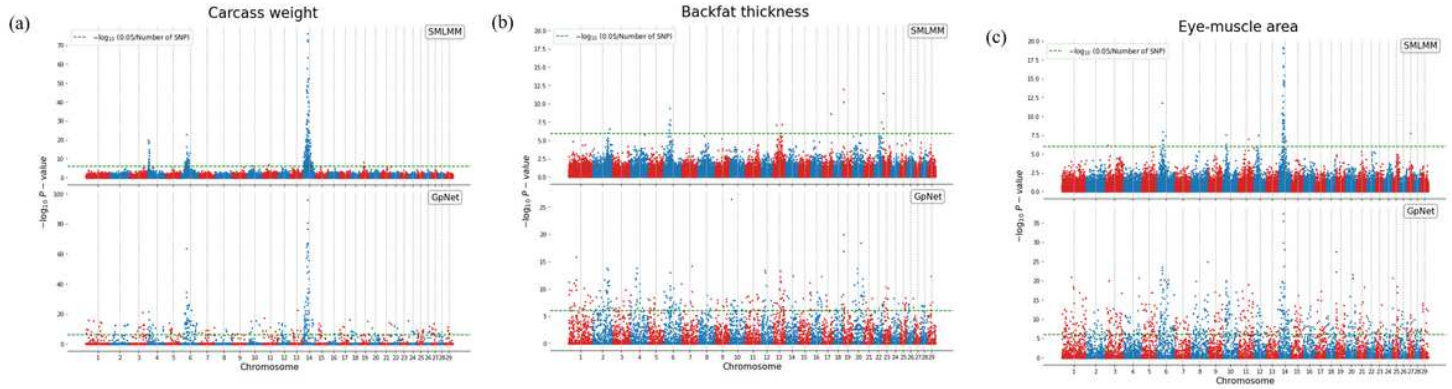

Figure 2

Manhattan plots of SMLMM and GpNet for each trait. (a) carcass weight; (b) backfat thickness; (c) eye-muscle area; x-axis is SNP position and chromosome, y-axis is − log10 P-value

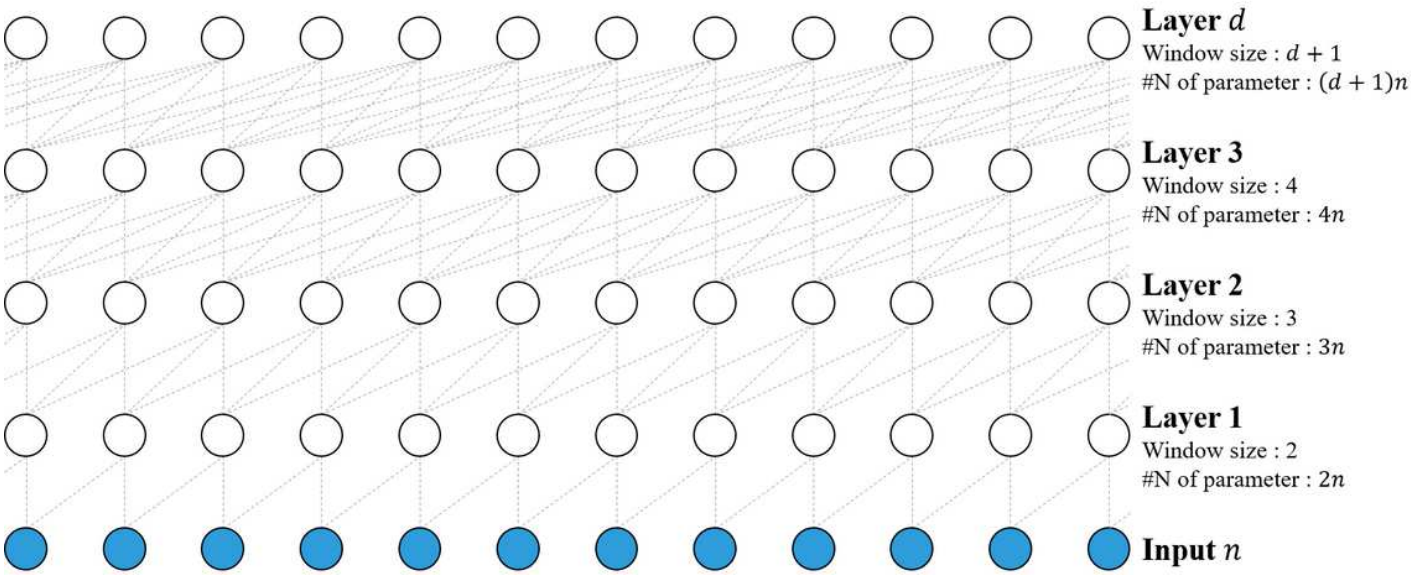

Figure 3

Visualization of the locally connected layer. n is the number of input SNPs, d means layer depths. Since weights are unshared at locally connected layer, the number of parameters at d-depth layer is (d + 1)n

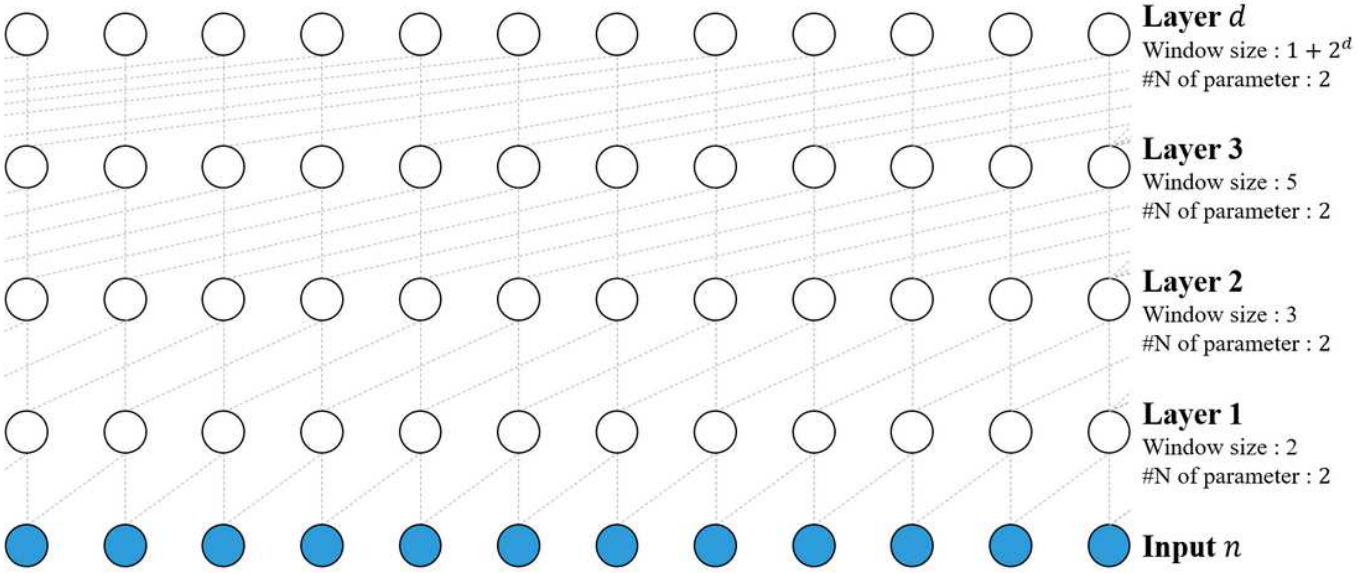

Figure 4

Visualization of dilated convolution. n is the number of input SNPs, d means layer depths. Since weights are shared at dilated convolution, the number of parameters at each layer is two.

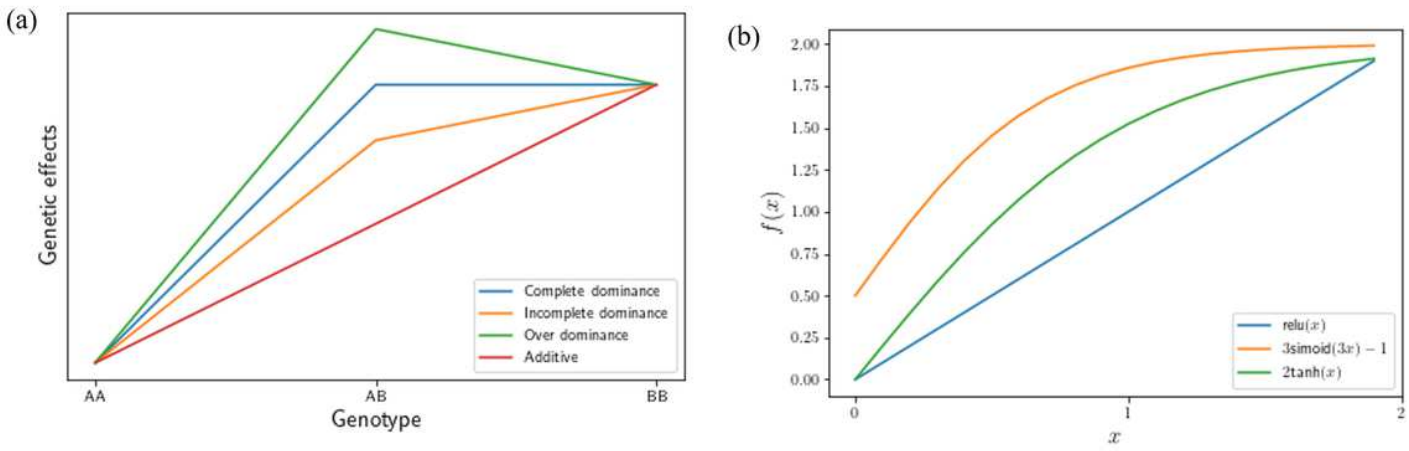

Figure 5

Type of dominance and candidate activations for dominance. (a) Type of dominance; (b) Candidate activations for dominance.

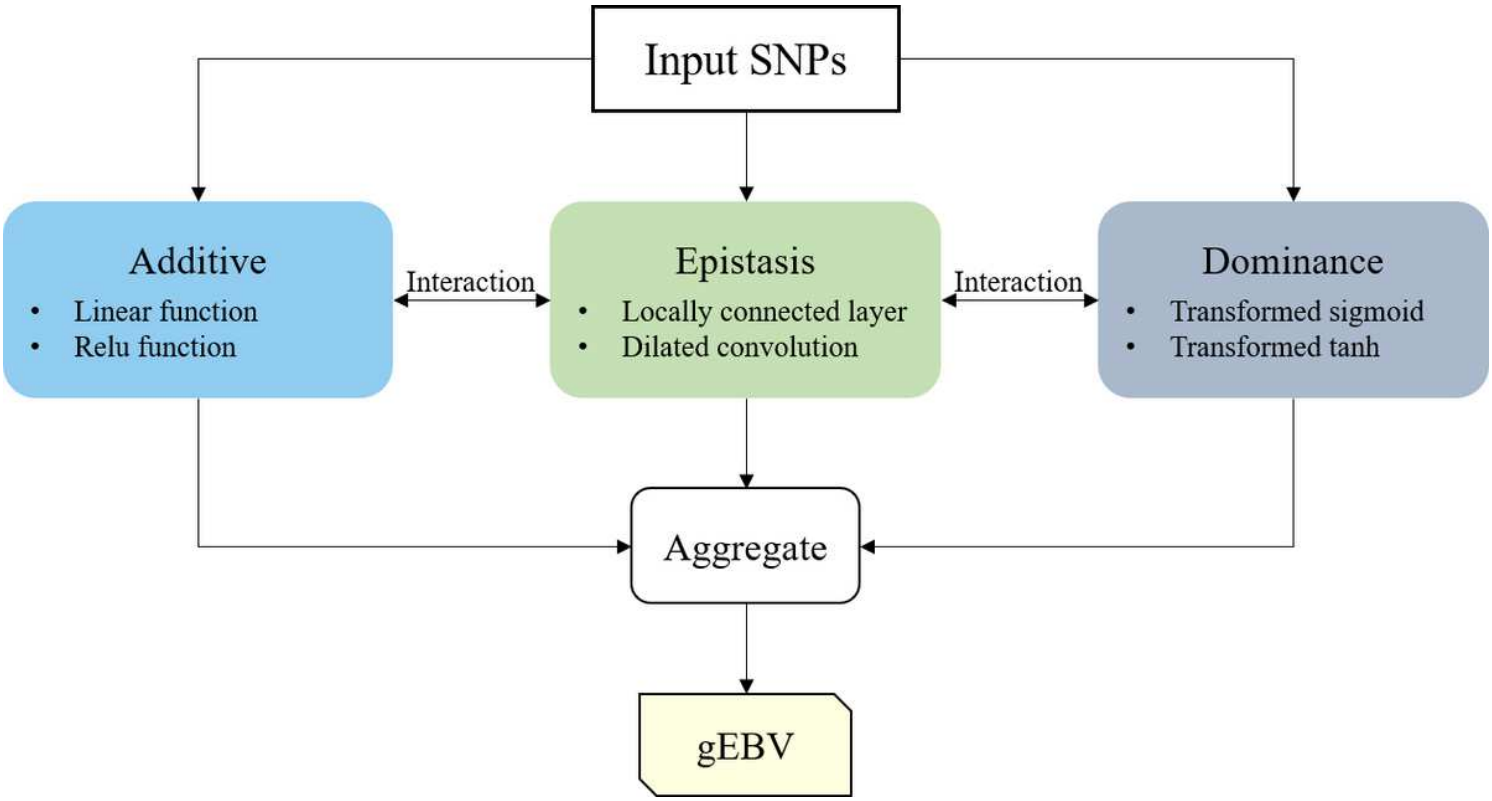

Figure 6

Identical model and candidate methods for different type of genetic effects.

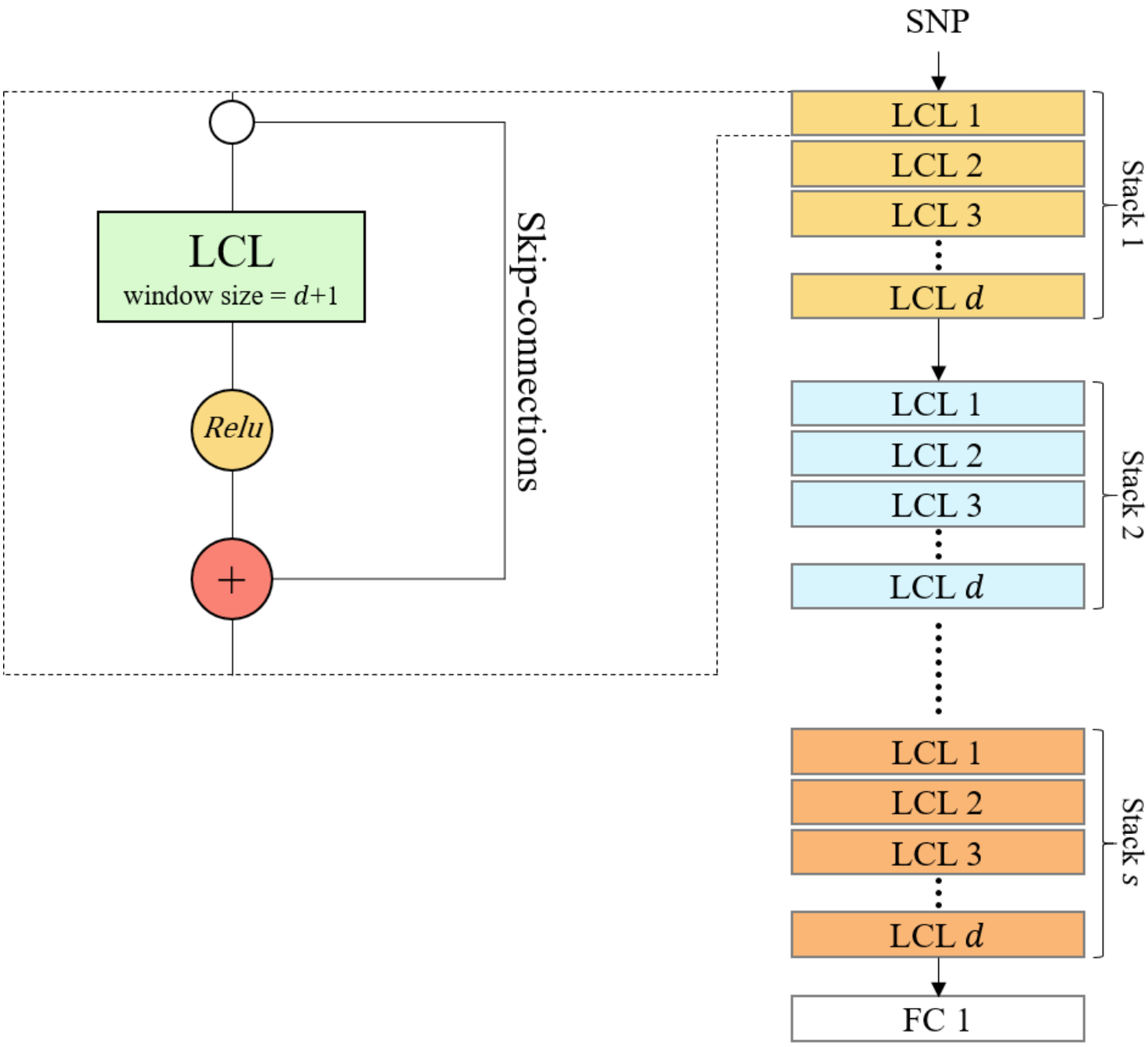

Figure 7

GpNet architecture. LCL is locally connected layer; d and s is the layer depths and the number of stacks; Relu is relu activation; GpNet can be scaled with d and s for different traits

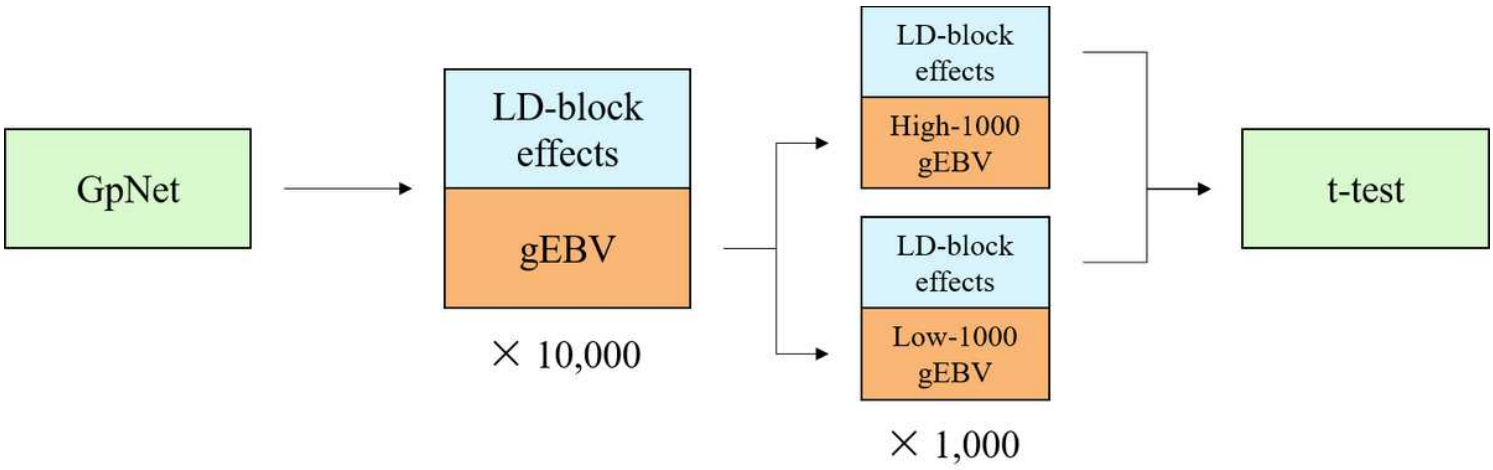

Figure 8

The process of QTL mapping.