Momen and Morota *Genet Sel Evol (2018) 50:45*
https://doi.org/10.1186/s12711-018-0415-9

Genetics Selection Evolution

SHORT COMMUNICATION

Open Access

Quantifying genomic connectedness and prediction accuracy from additive and non-additive gene actions

Mehdi Momen[1] and Gota Morota[2*]

Abstract

Background: Genetic connectedness is classically used as an indication of the risk associated with breeding value comparisons across management units because genetic evaluations based on best linear unbiased prediction rely for their success on sufficient linkage among different units. In the whole-genome prediction era, the concept of genetic connectedness can be extended to measure a connectedness level between reference and validation sets. However, little is known regarding (1) the impact of non-additive gene action on genomic connectedness measures and (2) the relationship between the estimated level of connectedness and prediction accuracy in the presence of non-additive genetic variation.

Results: We evaluated the extent to which non-additive kernel relationship matrices increase measures of connectedness and investigated its relationship with prediction accuracy in the cross-validation framework using best linear unbiased prediction and coefficients of determination. Simulated data assuming additive, dominance, and epistatic gene action scenarios and real swine data were analyzed. We found that the joint use of additive and non-additive genomic kernel relationship matrices or non-parametric relationship matrices led to increased capturing of connectedness, up to 25%, and improved prediction accuracies compared to those of baseline additive relationship counterparts in the presence of non-additive gene action.

Conclusions: Our findings showed that connectedness metrics can be extended to incorporate non-additive genetic variation of complex traits. Use of kernel relationship matrices designed to capture non-additive gene action increased measures of connectedness and improved whole-genome prediction accuracy, further broadening the scope of genomic connectedness studies.

Genetic connectedness is used to evaluate the extent to which reliable comparisons of estimated breeding values can be safely performed across management units. The strength of genetic links or connectedness relies on the relatedness of individuals across management units [1]. In turn, genetic evaluations of managed populations such as livestock species rely for their success on sufficient connectedness between different units. In such cases, best linear unbiased prediction (BLUP) provides fair ranking of the estimated breeding values of individuals while minimizing the risk of potential uncertainty in estimated breeding value comparisons [2–4]. The majority of previous studies on connectedness were performed with regard to pedigree relatedness; however, Yu et al. [5] rekindled an interest in this area by evaluating the utility of genome-based connectedness. Using real mice and cattle data, they reported that genomic relatedness enables the enhancement of genetic connectedness measures across management units compared to those obtained from pedigree relationships. This is mainly because genomic information captures relatedness between units that appears disconnected according to the pedigree. The utility of genomic connectedness was further investigated

*Correspondence: morota@vt.edu
[2] Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University, 175 West Campus Drive, Blacksburg, VA 24061, USA
Full list of author information is available at the end of the article

by assessing whether the enhanced estimates of connectedness delivered by genomics also led to an increased accuracy of breeding value prediction [6]. It was found that the use of genomic relatedness yields increased measures of connectedness and improved prediction accuracies (PA) compared to those of pedigree-based models under a purely additive gene action mode when a sufficient number of single-nucleotide polymorphisms (SNPs) is present. This parallels the recent recognition of the impact of non-additive genetic variation marked by SNPs e.g. [7, 8]. By properly accounting for non-additive genetic variation, it is potentially possible to enhance (1) the accuracy of total genetic value prediction, (2) the accuracy of breeding value prediction by clearly separating additive from non-additive genetic variation, and 3) the efficiency of mate allocation procedures as well as crossbreeding or purebred selection schemes [9, 10]. However, the relationship between the estimated level of connectedness and PA in the presence of non-additive genetic variation is less well understood. Accordingly, the objective of the current study was to evaluate the interrelationship between the degree of genomic connectedness and genome-enabled PA by calculating connectedness statistics from either the joint use of additive and non-additive genomic relationship matrices or non-parametric relationship matrices using simulated and real data, further broadening the scope of genomic connectedness studies.

Methods

Simulated data

A two-step simulation process was carried out using the QMSim software [11]. A historical population with 1000 individuals was created at the initial generation, followed by a sharp reduction in the population size owing to population bottleneck during generation 1 to 100. This resulted in the population size decreasing to 220 individuals in the last historical generation, creating initial linkage disequilibrium along with mutation and drift. The recent population was formed by randomly sampling 200 females and 10 males from the last historical generation. The individuals were mated for the subsequent five generations with equal probability of males and females, producing a total of 2000 individuals with a structured pedigree for analysis.

The simulated genome consisted of 29 pairs of autosomes each 100 cM long. To mimic a commercial Bovine 54K SNP chip, 1885 bi-allelic SNPs were equally distributed across each chromosome and each chromosome was assigned 65 quantitative trait loci (QTL). Phenotypes were simulated under three different gene action scenarios: (1) additive and dominance (AD), (2) additive, dominance, and epistasis (ADE), and (3) purely epistasis (PE). The simplest quantitative genetic model with main effects (additive and dominance) and epistasis constitutes a two-allele two-locus model. Epistasis was simulated only between pairs of QTL including second order additive × dominance (A×D) interactions. Five QTL from the 65 on each chromosome (total of 145) were selected to create 10,440 epistatic two-order interactions (145(145−1)/2 = 10,440). The total effect of QTL pairs influencing a given trait was calculated as the sum of all effects using the following model:

$$y_i = \sum_{k=1}^{nQTL} \mathbf{W}_{\mathbf{a}ik} a_k + \sum_{k=1}^{nQTL} \mathbf{W}_{\mathbf{d}ik} d_k + \sum_{k=1}^{nQTL} \sum_{k'=2}^{nQTL} \mathbf{l}_k \mathbf{l}_{k'} ad + \epsilon_i.$$

Here, a, d, and ad are the additive, dominance, and epistatic effects, respectively; $\mathbf{W_a}$, $\mathbf{W_d}$, and $\mathbf{l}_k\mathbf{l}_{k'}$ are SNP codes for additive, dominance, and epistasis, respectively; k denotes the kth QTL; and $nQTL$ is the number of QTL (for the epistatic term, this is only summed over the epistatic QTL). The phenotypic value of each individual y_i was created by adding a normally distributed residual $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ to the sum of genetic values. Additive effects were drawn from a Gamma distribution with shape and scale parameters equal to 0.42 and 8.282, respectively [12]. Their effect signs were sampled to be positive or negative with probability 0.5. The dominance effect for the kth QTL was determined as the product of the absolute value of the additive QTL effect and the degree of dominance $d_k = \delta_k \mid a_k \mid$ [13, 14]. Here, δ_k is the degree of dominance sampled from a normal distribution with $\delta_k \sim N(0, 1)$. The epistatic effects were drawn from a normal distribution with $N(0.02, \sigma^2 = 0.03)$ [14]. Additive and dominance components were simulated for the AD scenario; additive, dominance, and epistatic components were included for the ADE scenario; and only epistasis was considered for the PE scenario. Two broad-sense heritability levels (H^2) equal to 0.4 and 0.8 were simulated, with the partitioning of variance components shown in Table 1. We considered phenotypic variance equal to unity and simulated genetic variance according to the proportion of phenotypic variance explained by additive, dominance, and epistatic QTL effects: $\sigma_a^2 = \sum_k 2p_k q_k \alpha_k^2$, $\sigma_d^2 = \sum_k [2p_k q_k d_k]^2$, and $\sigma_{ad}^2 = 2 \sum_k \sum_{k'} p_k^2 p_{k'} q_{k'} (\alpha_k d_{k'})^2$, where $\alpha = [a + d(q-p)]^2$ is the allele substitution effect, and p and q are minor and major allele frequencies, respectively [15, 16].

Real data

For real data analysis, publicly available PIC swine data was used [17]. We analyzed five traits, T1, T2, T3, T4, and T5, with the corresponding number of individuals equal to 2804, 2715, 3141, 3184, and 3184. Their heritability values were 0.03, 0.23, 0.20, 0.32, and 0.36, respectively. It has been shown that this dataset exhibits a small to moderate amount of dominance genomic variation [18, 19]. Therefore, this dataset was considered suitable to test the extent to which the use of a non-additive genomic kernel relationship matrix might increase the capturing of connectedness measures. After removing SNPs with a minor allele frequency lower than 0.05, 52,842 SNPs remained for the analysis.

Management unit simulation

The management units were simulated according to the approach in Yu et al. [5] for simulated and real data. We clustered all individuals into management unit 1 (MU1) and management unit 2 (MU2) using the K-means clustering algorithm applied to a numerator relationship matrix computed from pedigree data such that the overall level of relatedness between individuals in different management units is minimized. There was no exchange of individuals between MU1 and MU2 in scenario 1 (S1), which served as a least connected design. An additional five management unit scenarios (S2 to S6) were considered by exchanging 10, 20, 30, 40, and 50% of individuals between MU1 and MU2 as shown in Fig. 1.

Genomic relationship kernel matrix

Three types of genomic relationship kernel matrices ($\mathbf{K}$) were used in the present study.

The additive genomic relationship matrix ($\mathbf{K} = \mathbf{G}$) was used to capture the pattern of additive inheritance $\mathbf{G} = \mathbf{W_a}\mathbf{W_a}'/2\sum_{k=1}^{m} p_k(1-p_k)$, where $\mathbf{W_a}$ is the centered marker incidence matrix taking values of $0 - 2p_k$ for zero copies of the reference allele, $1 - 2p_k$ for one copy of the reference allele, and $2 - 2p_k$ for two copies of the reference allele [20]. Here, p_k is the allele frequency at SNP $k = 1, \ldots, m$. The dominance genomic relationship matrix ($\mathbf{K} = \mathbf{D}$) aimed at capturing dominance gene action $\mathbf{D} = \mathbf{W_d}\mathbf{W_d}'/\sum_{k=1}^{m}(2p_k(1-p_k))^2$, where $\mathbf{W_d}$ is the dominance marker incidence matrix defined according to Vitezica et al. [21]. The additive by dominance genomic relationship matrix was constructed as $\mathbf{G}\#\mathbf{D}$, where # denotes the Hadamard product [22].

Gaussian kernel

The Gaussian kernel ($\mathbf{K} = \mathbf{GK}$) is equivalent to modeling epistatic gene action up to an infinite order by taking the Hadamard product between $\mathbf{G}$ matrices when SNPs were coded in an additive manner [23]. It is also known as a space continuous version of the diffusion kernel, which is deployed on graphs [24]. The Gaussian kernel between a pair of individuals i and j with their genotype vectors $\mathbf{w}_i \in (0,1,2)$ and $\mathbf{w}_j \in (0,1,2)$ is given by:

$$\begin{aligned}\mathbf{GK}(\mathbf{w}_i, \mathbf{w}_j) &= \exp(-\theta d_{ij}^2)\\ &= \prod_{k=1}^{m} \exp(-\theta(w_{ik} - w_{jk})^2),\end{aligned}$$

where $d_{ij} = \sqrt{(w_{i1} - w_{j1})^2 + \cdots + (w_{ik} - w_{jk})^2 + \cdots + (w_{im} - w_{jm})^2}$ is the Euclidean distance and θ is the smoothing parameter. Large θ leads to $\mathbf{GK}$ entries closer to 0 (i.e., local kernel) and smaller θ produces entries closer to 1 (i.e., global kernel). Therefore, θ controls the extent of genomic similarity between individuals.

Coefficient of determination

Consider a standard BLUP model, $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is the vector of the phenotypes, $\mathbf{X}$ and $\mathbf{Z}$ are the incidence matrices for systematic and random effects, respectively, $\mathbf{b}$ and $\mathbf{u}$ are the vectors of systematic effects and genetic values, and $\boldsymbol{\epsilon}$ is the vector of residuals. By defining $\mathrm{var}(\mathbf{u}) = \mathbf{K}\sigma_u^2$ we have:

$$\begin{aligned}\mathrm{BLUP}(u) &= \sigma_u^2\mathbf{K}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})\\ &= \sigma_u^2\mathbf{K}\mathbf{Z}'\mathbf{P}\mathbf{y}\\ \mathrm{var}(\hat{\mathbf{u}}) &= \sigma_u^2\mathbf{K}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{K}\sigma_u^2\end{aligned}$$

where σ_u^2 is the variance associated with a kernel matrix $\mathbf{K}$, $\mathbf{V}$ is the variance of $\mathbf{y}$, and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ [25]. Recall that since $\mathrm{cov}(\hat{\mathbf{u}}, \mathbf{u}') = \mathrm{cov}(\mathbf{u}', \hat{\mathbf{u}}) = \mathrm{var}(\hat{\mathbf{u}})$, the prediction error variance (PEV) of $\mathbf{u}$ is given by:

$$\begin{aligned}\mathrm{PEV} &= \mathrm{var}(\hat{\mathbf{u}} - \mathbf{u})\\ &= \mathrm{var}(\hat{\mathbf{u}}) + \mathrm{var}(\mathbf{u}) - 2\mathrm{cov}(\hat{\mathbf{u}}, \mathbf{u}')\\ &= \mathrm{var}(\mathbf{u}) - \mathrm{var}(\hat{\mathbf{u}})\\ &= \mathbf{K}\sigma_u^2 - \sigma_u^2\mathbf{K}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{K}\sigma_u^2,\end{aligned}$$

Table 1 Simulated heritability value for each gene action scenario

H^2	Gene action	h_A^2	h_D^2	h_E^2
0.4	AD	0.3	0.1	-
	ADE	0.2	0.1	0.1
	PE	–	–	0.4
0.8	AD	0.6	0.2	-
	ADE	0.4	0.2	0.2
	PE	–	–	0.8

H^2, h_A^2, h_D^2, and h_E^2 are broad-sense, additive, dominance, and epistatic heritabilities, respectively. Gene action scenarios AD, ADE, and PE denote additive and dominance, additive, dominance, and epistasis, and purely epistasis, respectively

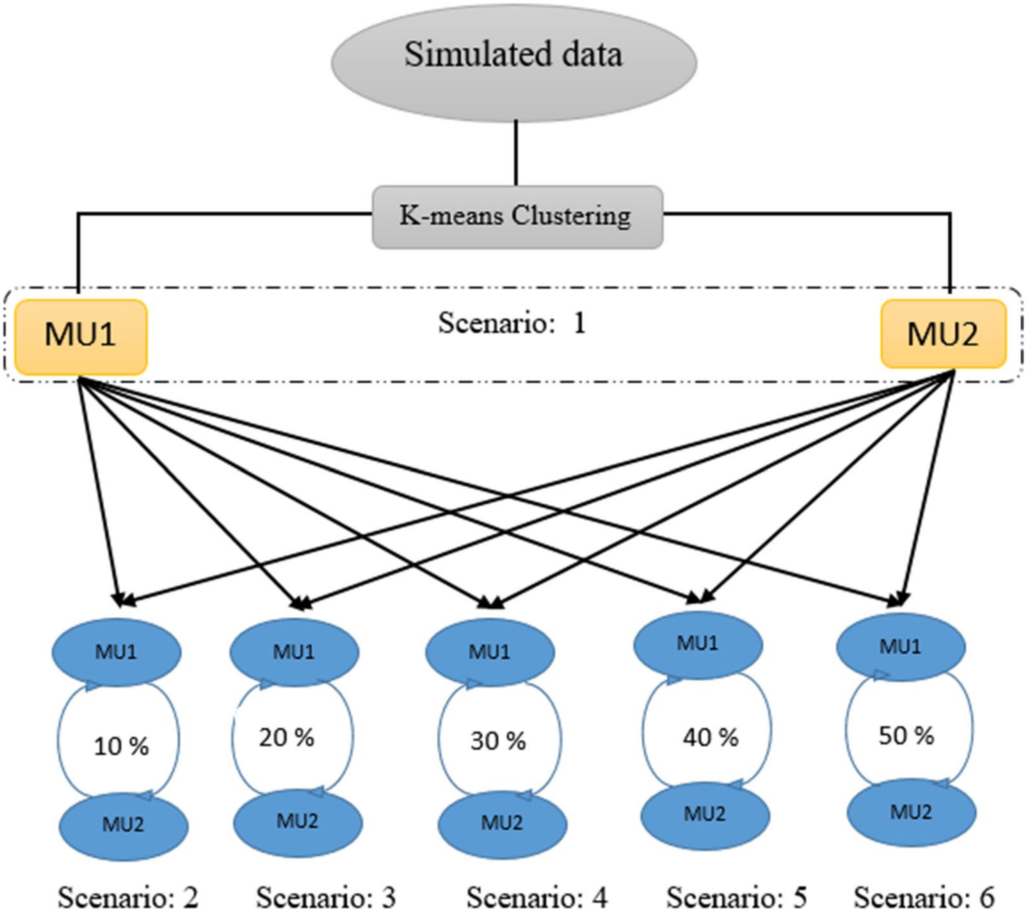

Fig. 1 Simulated management units (MU). Scenario 1: Disconnected management units MU1 and MU2. Scenario 2: 10% of individuals were exchanged between MU1 and MU2. Scenario 3: 20% of individuals were exchanged between MU1 and MU2. Scenario 4: 30% of individuals were exchanged between MU1 and MU2. Scenario 5: 40% of individuals were exchanged between MU1 and MU2. Scenario 6: 50% of individuals were exchanged between MU1 and MU2

where **K** can be any positive (semi)definite relationship matrix between pairs of individuals discussed earlier.

The generalized coefficient of determination (CD), which is also known as the square of the correlation between the predicted and the true difference in the genetic values, was used to quantify connectedness. CD of the contrast between management units l and l' consisting of n_l and $n_{l'}$ individuals is given by [26, 27]:

$$\begin{aligned} \mathrm{CD} &= \frac{\mathbf{x}_{ll'}\mathrm{var}(\hat{\mathbf{u}})\mathbf{x}_{ll'}}{\mathbf{x}_{ll'}\mathrm{var}(\mathbf{u})\mathbf{x}_{ll'}} \\ &= \frac{\mathbf{x}_{ll'}[\mathrm{var}(\mathbf{u}) - \mathrm{var}(\hat{\mathbf{u}} - \mathbf{u})]\mathbf{x}_{ll'}}{\mathbf{x}_{ll'}\mathrm{var}(\mathbf{u})\mathbf{x}_{ll'}} \\ &= 1 - \frac{\mathbf{x}_{ll'}[\mathrm{var}(\hat{\mathbf{u}} - \mathbf{u})]\mathbf{x}_{ll'}}{\mathbf{x}_{ll'}\mathrm{var}(\mathbf{u})\mathbf{x}_{ll'}}, \end{aligned}$$

where $\mathbf{x}$ is the contrast vector involving $1/n_l$, $-1/n_{l'}$ and 0 corresponding to individuals belonging to lth, l'th, and the remaining units. Here, the sum of contrast vector elements is zero. The greater the CD of contrast, the greater the connectedness. A large CD is expected when prediction error covariance in the numerator is large, reflecting errors that are in the same direction between units. Alternatively, the measure of CD decreases when the relationship between individuals across units is large in the denominator. Therefore, the CD of contrast combines the prediction error variance of the difference (PEVD) [2] and genetic variability. This metric was chosen because it was found to represent the most stable connectedness metric in a recent study [5].

Connectedness measures and prediction accuracy

Measures of CD between MU1 and MU2 were inferred from estimated variance components followed by assessing genomic PA by two-fold cross-validation using a BLUP type model. In the first fold, MU1 was treated as a training set and MU2 was treated as a testing set. This was reversed in the second fold such that MU2 was used to train the model and MU1 was used to test prediction performance. The multi-kernel **G** and **D** approach in the AD scenario, the multi-kernel **G**, **D**, and **G#D** approach in the ADE scenario, and the **GK** matrix in the PE scenario were benchmarked against the baseline **G** matrix (i.e., genomic BLUP). Note that the use of **GK** corresponds to fitting a reproducing kernel Hilbert spaces regression (e.g. [28]). For a multi-kernel approach, we weighted each kernel by its relative contribution to the marked total genetic variation, also known as kernel averaging or multiple kernel learning [29], to measure connectedness and assess PA. For instance, the kernel matrix $\mathbf{K} = \frac{\sigma_g^2}{\sigma_g^2+\sigma_d^2}\mathbf{G} + \frac{\sigma_d^2}{\sigma_g^2+\sigma_d^2}\mathbf{D}$ was used when **G** and **D** were fitted together, where σ_g^2 and σ_d^2 were additive and dominance genomic variances, respectively. PA was obtained as the correlation between true and predicted genetic values for the simulated data averaged across 10 replicates ($\mathrm{cor}(\mathbf{g}, \hat{\mathbf{g}})$) and the correlation between phenotypes and predicted genetic values for the real data ($\mathrm{cor}(\mathbf{y}, \hat{\mathbf{g}})$).

Results

AD scenario

The relationships between CD and PA across the six management unit simulation scenarios (S1 to S6) are shown in Fig. 2. The joint fit of **G** and **D** kernel relationship matrices was benchmarked using the **G** matrix alone. A sharp increase in PA was observed with the increasing proportion of exchanged individuals from S1 to S3, which reached a plateau after 30% exchange rate between MU1 and MU2 in S4. Overall, PA improved as more individuals between MU1 and MU2 were shared. Higher PA values were achieved by accounting for dominance **G** + **D** compared to **G** alone for the two heritability levels considered (0.4 and 0.80). The lowest PA (0.368) was obtained in S1 with **G** and the highest PA (0.632) was obtained in S4 with **G** + **D**.

For the measures of connectedness, there was a good agreement between increasing the rate of exchange and stronger measures of connectedness up to S3. However, the estimates of CD increased up to scenario S3, followed by a decrease from scenario S4 onward because CD penalizes connectedness measures when two units are genetically close. The results showed that establishing genetic links between management units by exchanging more individuals created more genetic similarity on one side and reduced genetic variability on the other side, resulting in lower CD values. CD of contrast measured by **G** + **D** captured stronger connectedness than that of **G** consistently across all scenarios (S1 to S6). The largest measured CD (0.989) was obtained with **G** + **D** in S3, and the smallest CD (0.64) was obtained with **G** in S6. Overall, accounting for dominance variation increased PA and measures of CD. The relationship between PA and CD was positively associated up to S3; then, whereas PA continued to increase, CD began to level off.

ADE scenario

The results of PA and CD from the ADE scenario are shown in Fig. 3. We found that the overall pattern resembled that of the AD scenario. That is, with increasing degree of similarity among management units, PA increased and then reached a plateau after S4. The highest PA (0.731) was obtained with **G** + **D** + **G#D** kernel matrices in S4 and the smallest PA (0.245) with **G** in S1.

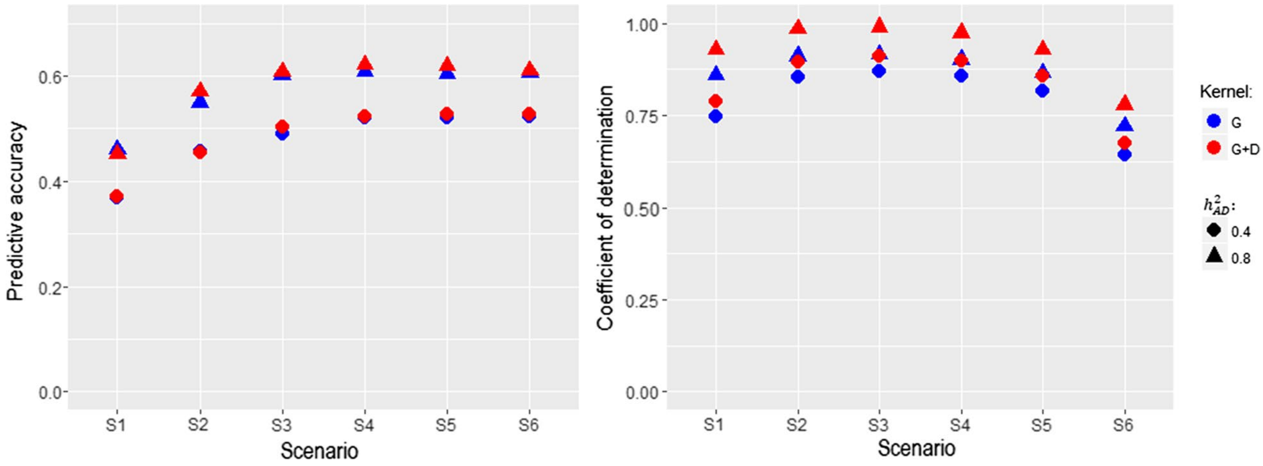

Fig. 2 Relationship between prediction accuracies (left panel) and connectedness measures (right panel) under an additive and dominance scenario. The magnitude of the relationship level was steadily increased from scenario 1 (S1) to scenario 6 (S6). **G**: additive genomic kernel relationship matrix. **D**: dominance genomic kernel relationship matrix. h^2_{AD}: broad-sense heritability including additive and dominance variation

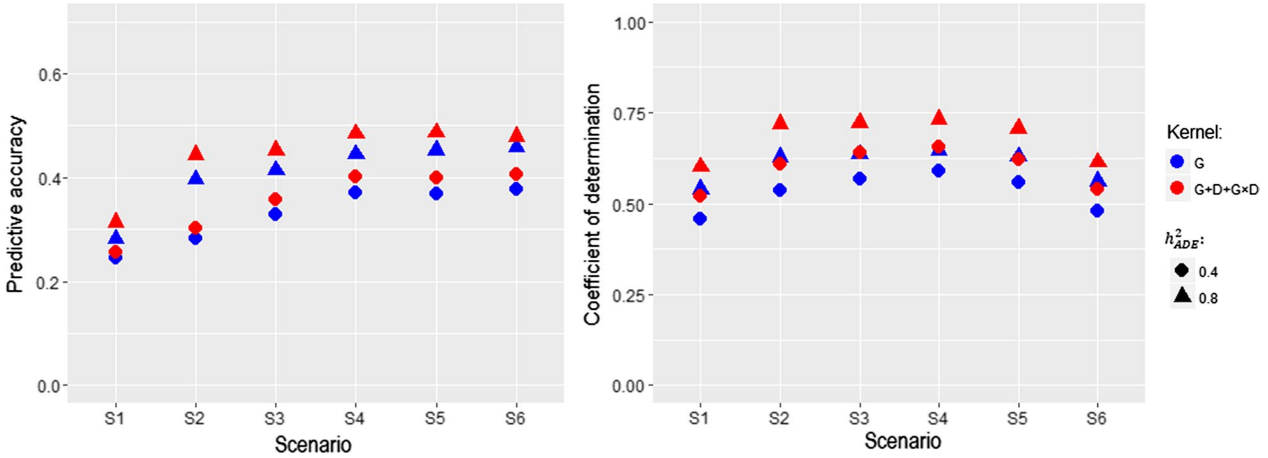

Fig. 3 Relationship between prediction accuracies (left panel) and connectedness measures (right panel) under an additive, dominance, and epistasis scenario. The magnitude of the relationship level was steadily increased from scenario 1 (S1) to scenario 6 (S6). **G**: additive genomic kernel relationship matrix. **D**: dominance genomic kernel relationship matrix. **G** × **D**: additive × dominance genomic kernel relationship matrix. h^2_{ADE}: broad-sense heritability including additive, dominance, and epistatic variation

The PA results suggested that increasing the number of linking individuals improves PA and the use of non-additive genomic relationship matrices simultaneously further increased PA.

In comparison, measures of CD were strengthened with the increase of linking individuals up to S4, followed a decreasing tendency, similar to the pattern observed in the AD scenario. Improved capture of connectedness was achieved by explicitly accounting for additive, dominance, and epistasis variations compared to additive only. The greatest and weakest measures of connectedness were observed with $\mathbf{G} + \mathbf{D} + \mathbf{G\#D}$ (0.731) in S4 and with **G** (0.456) in S1, respectively.

PE scenario

Performance of **GK** and **G** was compared in the PE scenario. We considered different values for the smoothness parameter θ ranging from 0.22, 0.5, and 0.9 to 1.6. These θ values were chosen such that the averages of off-diagonal elements corresponded to 0.8, 0.6, 0.4, and 0.2 covering global to local kernels (Fig. 4). The relationship between PA and CD for **GK** and **G** is shown in Fig. 5. For $H^2 = 0.4$, the results from the PE scenario were similar to those of AD and ADE scenarios, showing a higher PA with an increasing number of linking individuals. A θ equal to 1.6 produced the highest overall PA. Altogether, these results demonstrate the usefulness of **GK** to capture information arising from non-additive genetic variation. The advantage of **GK** over **G** for PA was less obvious when heritability was high ($H^2 = 0.8$).

The right side panel in Fig. 5 illustrates how θ impacts the measures of connectedness under the PE scenario. For $H^2 = 0.40$, the largest CD value was obtained in S3 with $\mathbf{GK}(\theta) = 1.6$, and the smallest values were observed in S1 and S6 with $\mathbf{GK}(\theta) = 0.22$. The connectedness measures from **G** were between these two extreme **GK**. Again, the highest PA was observed in S6 whereas the highest CD was observed in S3. This is because CD accounts for the reduction of connectedness owing to low genetic diversity [27]. A similar pattern was observed for $H^2 = 0.8$, highlighting that the utility of **GK** to capture connectedness under non-additive gene actions also holds for a highly heritable trait.

Real data

The results from real data are shown in Fig. 6. As more individuals between the two units were exchanged, PA increased across all traits until a maximum was reached whereas CD started to drop in S5. Fitting **G** and **D** simultaneously yielded better prediction in almost all cases and also captured greater amounts of connectedness than those of **G** alone. Traits with a higher heritability (e.g. T4 and T5) presented higher PA and greater CD levels than those with a lower heritability (e.g. T1). The results from real data analysis corroborated the utility of the multi-kernel approach from the simulation study.

Discussion

The assessment of genetic connectedness originated from testing the estimability of linear functions of fixed effects in n-way cross classifications to determine the absence or presence of connectedness [30, 31]. It was subsequently extended to the random effects framework [1] to quantify

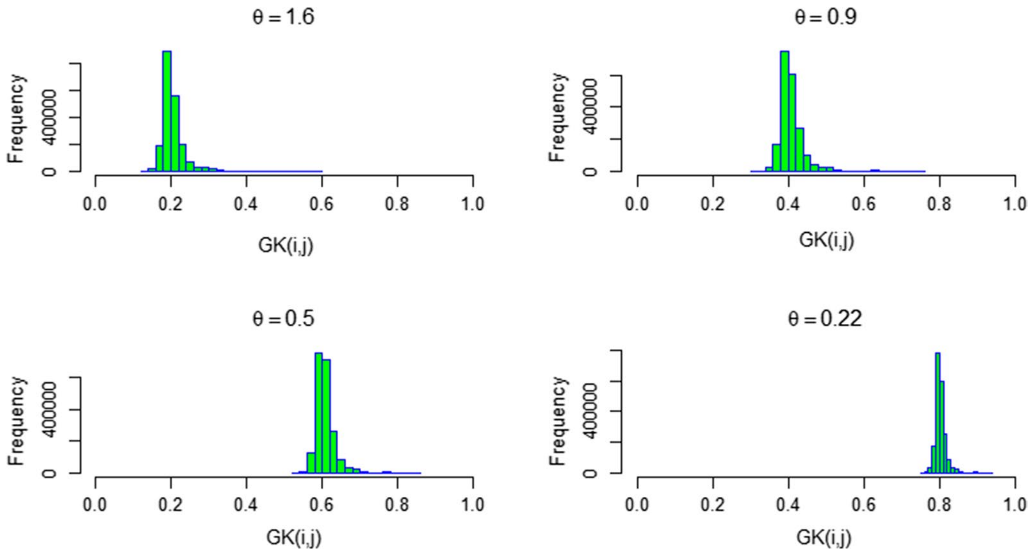

Fig. 4 Histogram of off-diagonal elements between individual *i* and *j* for the Gaussian kernel matrix **GK**(i, j) with different smoothness parameters $\theta = 1.6, 0.9, 0.5$, and 0.22

the uncertainty associated with the accuracy of breeding value comparisons involving different management units. In this sense, connectedness is a measure germane to the capability to have estimable comparisons [3]. In the genomics era, the concept of genetic connectedness offers insights on two aspects of the prediction of genetic values. The first is relevant to improving the quality of genomic breeding value comparisons [5, 32] whereas the other is related to improving the accuracy of genomic prediction [33]. Notably, it is possible to reconcile these two items by quantifying a genomic connectedness level between reference and validation sets in the whole-genome prediction paradigm. Toward this end, Yu et al. [6] investigated the relationship between connectedness measures and PA using pedigree and genomic information under an additive model.

Concurrently, it has been shown that whole-genome prediction models designed to capture non-additivity yield slightly to moderately higher PA than additive counterparts when the underlying genetic architecture is governed by dominance or epistasis e.g. [28, 34]. Although the extent of non-additive genetic variance may not be big in general, this type of variance is particularly important for fitness-related traits [35]. These recent findings served as the impetus for the present study, extending the scope of connectedness applications by further considering non-additive genetic variation.

We observed that the inclusion of non-additive genetic relationship kernel matrices or non-parametric relationship matrices in a BLUP type model increased PA as more individuals were exchanged between MU1 and MU2, and that this was associated with stronger measures of connectedness up to S3 or S4. This reinforced the view that the commonly observed higher prediction performance in non-additive or non-parametric models in the presence of non-linear gene action is due to improved capturing of connectedness between units. We also found that the choice of smoothness parameter θ not only influences PA but also the extent of CD. This indicates the importance of the smoothness parameter in evaluating PA and CD, especially when a complex trait is controlled by non-additive gene actions. In general, our results showed that when the optimum θ is selected, PA and CD of **GK** will be better than those of **G**, and that even **GK** constructed from additive coding of SNPs only captures additive by additive epistasis theoretically [23]. We note that many studies have shown that PA decreases when the reference population has a lower relatedness to the validation population e.g. [36, 37]. This is equivalent to when two units exhibit weak connectedness. Use of connectedness thereby opens up the possibility for an alternative way to measure the strength of relationship between these two populations instead of using an average relationship.

Moreover, once the rate of exchange reached S3 or S4, the estimated level of CD gradually leveled off in all management unit simulation scenarios, contrary to PA. This is because when there are sufficient numbers of individuals linking MU1 and MU2, the denominator of CD becomes smaller thus increasing the second term, which in turn renders the CD of contrast to become small. This

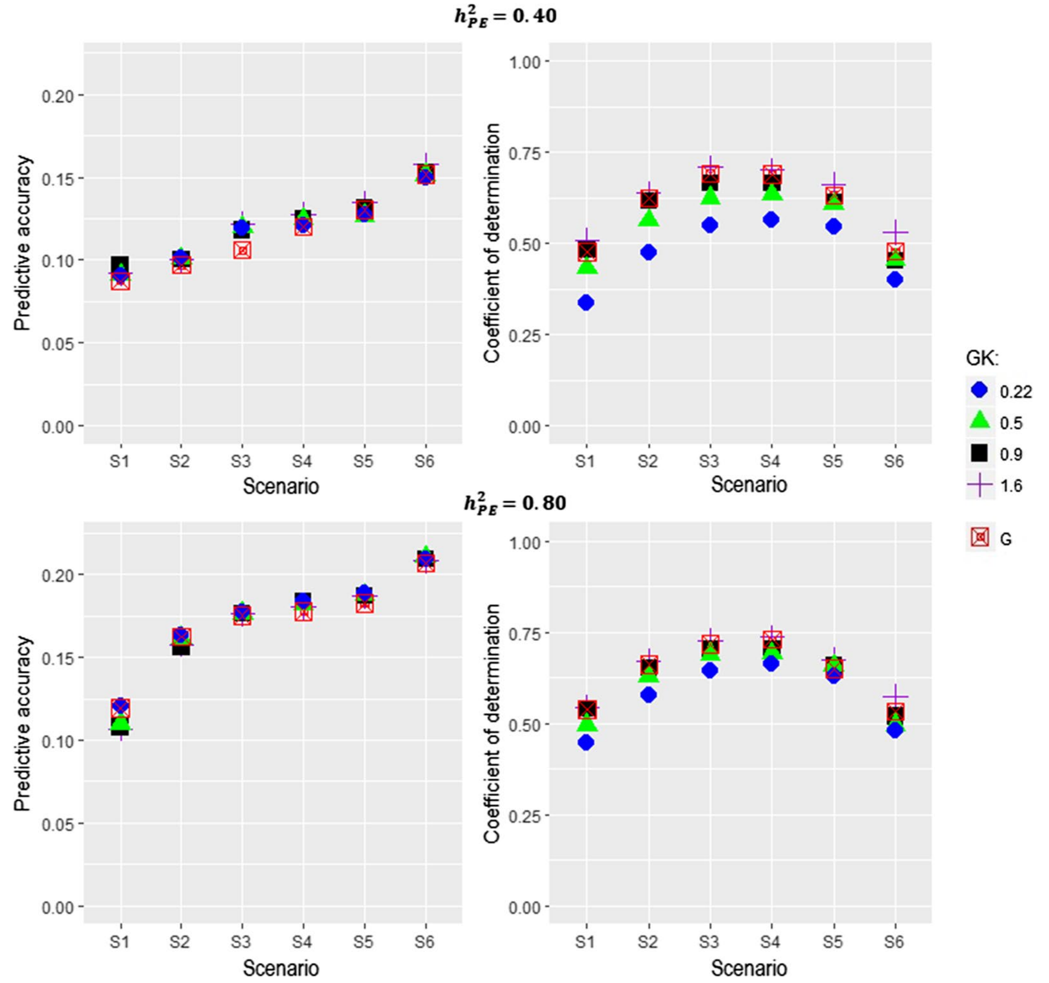

Fig. 5 Relationship between prediction accuracies (left panel) and connectedness measures (right panel) under a purely epistasis scenario. The magnitude of the relationship level was steadily increased from scenario 1 (S1) to scenario 6 (S6). **GK**: Gaussian kernel relationship matrix with the smoothness parameters $\theta = 1.6, 0.9, 0.5$, and 0.22. **G**: additive genomic kernel relationship matrix. h^2_{PE}: broad-sense heritability including epistatic variation

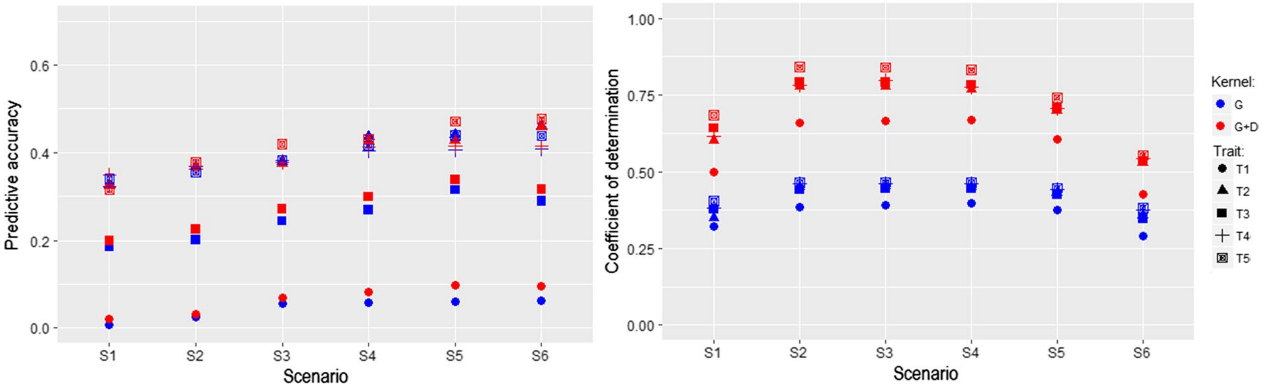

Fig. 6 Relationship between prediction accuracies (left panel) and connectedness measures (right panel) in the real swine data. The magnitude of the relationship level was steadily increased from scenario 1 (S1) to scenario 6 (S6). **G**: additive genomic kernel relationship matrix. **D**: dominance genomic kernel relationship matrix. T1 to T5 denote five different traits analyzed in this study

agrees with the findings in other studies dealing with only additive genetic variation [5, 6]. Together, these findings suggest that the use of CD holds great potential to identify an optimal breeding program design in terms of genetic diversity while maximizing PA, whereas other connectedness metrics such as PEVD aim at increasing PA regardless of how closely individuals between units become related [5]. Note that PA is one of the criteria to determine the most appropriate model to fit (for example, **G** vs. **G** + **D**). Once the model is chosen, CD can be used to identify an appropriate level of relatedness or diversity between two units while maintaining high PA.

Although we applied *K*-means clustering of a numerator relationship matrix, the choice of **K** for clustering may impact our results. Thus, we further constructed management units based on clustering of **G** or **G** + **D** under the AD scenario. As shown in Figures S1 and S2 (see Additional file 1: Figures S1 and S2), *K*-means clustering of **G** or **G** + **D** produced patterns of PA and CD that are similar to those generated using the numerator relationship. We also repeated our analyses using forward validation rather than *K*-means clustering. We treated 1200 individuals in generations 1 to 3 as the training set (MU1) and 800 individuals in generations 4 to 5 as the testing set (MU2) under the AD scenario. We found that using **G** + **D** yielded higher PA and greater amount of CD compared to using **G** (Additional file 1: Figure S3).

The utility of genomic connectedness does not preclude its application in management units. For instance, connectedness measured by CD is currently gaining recognition for training population formation in plant breeding [38]. We contend that the use of CD holds promise to tackle a multitude of challenges related to increasing genomic prediction while maintaining genetic diversity.

Conclusion

Here, the genetic connectedness metric, CD, was used to assess genomic connectedness measures between reference and validation sets in a whole-genome prediction framework using simulated and real data in the presence of non-additive gene action. Joint fitting of additive and non-additive genomic kernel relationship matrices or non-parametric relationship matrices could yield enhanced capture of connectedness and improved PA compared to those obtained through baseline additive models. Our approach shows promise to measure connectedness levels and investigate their relationship with genomic PA when the linear assumption of genotype-phenotype mapping may not hold.

Additional file

Additional file 1: Figure S1. Relationship between prediction accuracies (left panel) and connectedness measures (right panel) under an additive and dominance scenario based on the *K*-means clustering using the genomic relationship matrix. The magnitude of the relationship level was steadily increased from scenario 1 (S1) to scenario 6 (S6). $\mathbf{G}$: additive genomic kernel relationship matrix. $\mathbf{D}$: dominance genomic kernel relationship matrix. h^2_{AD}: broad-sense heritability including additive and dominance variation. **Figure S2.** Relationship between prediction accuracies (left panel) and connectedness measures (right panel) under an additive and dominance scenario based on the *K*-means clustering using the multikernel genomic and dominance relationship matrix. The magnitude of the relationship level was steadily increased from scenario 1 (S1) to scenario 6 (S6). $\mathbf{G}$: additive genomic kernel relationship matrix. $\mathbf{D}$: dominance genomic kernel relationship matrix. h^2_{AD}: broad-sense heritability including additive and dominance variation. **Figure S3.** Relationship between prediction accuracies (left panel) and connectedness measures (right panel) under an additive and dominance scenario based on forward validation. $\mathbf{G}$: additive genomic kernel relationship matrix. $\mathbf{D}$: dominance genomic kernel relationship matrix. h^2: heritability.

Authors' contributions

MM performed analyses, interpreted the results, and drafted the manuscript. GM conceived the study, interpreted the results, and revised the manuscript. Both authors read and approved the final manuscript.

Author details

[1] Department of Animal Science, Faculty of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran. [2] Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University, 175 West Campus Drive, Blacksburg, VA 24061, USA.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

All information supporting the results is included in the text, figures and tables of this article. The publicly available swine dataset can be downloaded from Cleveland et al. [17].

Ethics approval and consent to participate

Not applicable.

Funding

GM acknowledges funding from the University of Nebraska and Virginia Polytechnic Institute and State University startup funds.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 8 May 2018 Accepted: 5 September 2018
Published online: 17 September 2018

References

1. Foulley J, Bouix J, Goffinet B. Connectedness in genetic evaluation. In: Gianola D, Hammond K, editors. Advances in statistical methods for genetic improvement of livestock. Berlin: Springer; 1990. p. 277–308.
2. Kennedy BW, Trus D. Considerations on genetic connectedness between management units under an animal model. J Anim Sci. 1993;71:2341–52.

3. Mathur PK, Sullivan BP, Chesnais JP. Measuring connectedness: concept and application to a large industry breeding program. In: Proceedings of 7th world congress on genetics applied to livestock production, Montpellier, 19–23 August 2002. 2002.
4. Kuehn LA, Notter DR, Nieuwhof GJ, Lewis RM. Changes in connectedness over time in alternative sheep sire referencing schemes. J Anim Sci. 2008;86:536–44.
5. Yu H, Spangler ML, Lewis RM, Morota G. Genomic relatedness strengthens genetic connectedness across management units. G3 (Bethesda). 2017;7:3543–56.
6. Yu H, Spangler ML, Lewis RM, Morota G. Stronger measures of genomic connectedness enhance prediction accuracies across management units. In: Proceedings of the 11th world congress on genetics applied to livestock production, Auckland, 11–16 February 2018. 2018. p. 406.
7. Esfandyari H, Bijma P, Henryon M, Christensen OF, Sørensen AC. Genomic prediction of crossbred performance based on purebred Landrace and Yorkshire data using a dominance model. Genet Sel Evol. 2016;48:40.
8. Forneris NS, Vitezica ZG, Legarra A, Pérez-Enciso M. Influence of epistasis on response to genomic selection using complete sequence data. Genet Sel Evol. 2017;49:66.
9. Aliloo H, Pryce JE, González-Recio O, Cocks BG, Hayes BJ. Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. Genet Sel Evol. 2016;48:8.
10. Varona L, Legarra A, Toro MA, Vitezica ZG. Non-additive effects in genomic selection. Front Genet. 2018;9:78.
11. Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 2009;25:680–1.
12. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
13. Wellmann R, Bennewitz J. The contribution of dominance to the understanding of quantitative genetic variation. Genet Res (Camb). 2011;93:139–54.
14. Wittenburg D, Melzer N, Reinsch N. Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers. BMC Genet. 2011;12:74.
15. Cheverud JM, Routman EJ. Epistasis and its contribution to genetic variance components. Genetics. 1995;139:1455–61.
16. Holland JB. Epistasis and plant breeding. Plant Breed Rev. 2001;21:27–92.
17. Cleveland MA, Hickey JM, Forni S. A common dataset for genomic analysis of livestock populations. G3 (Bethesda). 2012;2:429–35.
18. 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. PLoS One. 2014;9:e87666.
19. Nishio M, Satoh M. Including dominance effects in the genomic BLUP method for genomic evaluation. PLoS One. 2014;9:e85792.
20. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
21. Vitezica ZG, Varona L, Legarra A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics. 2013;195:1223–30.
22. Henderson CR. Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. J Anim Sci. 1985;60:111–7.
23. Jiang Y, Reif JC. Modeling epistasis in genomic selection. Genetics. 2015;201:759–68.
24. Morota G, Koyama M, Rosa GJ, Weigel KA, Gianola D. Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. Genet Sel Evol. 2013;45:17.
25. Henderson CR. Applications of linear models in animal breeding. In: Schaeffer LR, editor. 3rd ed. Guelph: University of Guelph; 1984.
26. Laloë D. Precision and information in linear models of genetic evaluation. Genet Sel Evol. 1993;25:557–576.
27. Laloë D, Phocas F, Ménissier F. Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. Genet Sel Evol. 1996;28:359–78.
28. Morota G, Gianola D. Kernel-based whole-genome prediction of complex traits: a review. Front Genet. 2014;5:363.
29. de los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res (Camb). 2010;92:295–308.
30. Petersen PH. A test for connectedness fitted for the two-way BLUP-sire evaluation. Acta Agric Scand. 1978;28:360–2.
31. Fernando RL, Gianola D, Grossman M. Identifying all connected subsets in a two-way classification without interaction. J Dairy Sci. 1983;66:1399–402.
32. Holmes JB, Dodds KG, Lee MA. Estimation of genetic connectedness diagnostics based on prediction errors without the prediction error variance-covariance matrix. Genet Sel Evol. 2017;49:29.
33. Habier D, Fernando RL, Dekkers JC. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177:2389–97.
34. Howard R, Carriquiry AL, Beavis WD. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 (Bethesda). 2014;4:1027–46.
35. Crow JF. On epistasis: why it is unimportant in polygenic directional selection. Philos Trans R Soc Lond B Biol Sci. 2010;365:1241–4.
36. Pszczola M, Strabel T, Mulder HA, Calus MP. Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci. 2012;95:389–400.
37. Clark SA, Hickey JM, Daetwyler HD, van der Werf JH. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol. 2012;44:4.
38. Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, et al. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics. 2012;192:715–28.