bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

GCA: An R package for genetic connectedness analysis using pedigree and genomic data

Haipeng Yu[1*] and Gota Morota[1*]

[1]Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA 24061

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Running title: Genetic connectedness analysis software

ORCID: 0000-0002-8923-9733 (HY) and 0000-0002-3567-6911 (GM).

* Corresponding author:

Haipeng Yu
Department of Animal and Poultry Sciences
Virginia Polytechnic Institute and State University
175 West Campus Drive
Blacksburg, Virginia 24061 USA.
E-mail: haipengyu@vt.edu

Gota Morota
Department of Animal and Poultry Sciences
Virginia Polytechnic Institute and State University
175 West Campus Drive
Blacksburg, Virginia 24061 USA.
E-mail: morota@vt.edu

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Background: Genetic connectedness is a critical component of genetic evaluation as it assesses the comparability of predicted genetic values across units. Genetic connectedness also plays an essential role in quantifying the linkage between reference and validation sets in whole-genome prediction. Despite its importance, there is no user-friendly software tool available to calculate connectedness statistics.

Results: We developed the GCA R package to perform genetic connectedness analysis for pedigree and genomic data. The software implements a large collection of various connectedness statistics as a function of prediction error variance or variance of unit effect estimates. The GCA R package is available at GitHub and the source code is provided as open source.

Conclusions: The GCA R package allows users to easily assess the connectedness of their data. It is also useful to determine the potential risk of comparing predicted genetic values of individuals across units or measure the connectedness level between training and testing sets in genomic prediction.

Keywords: genetic connectedness, prediction error of variance, variance of unit effect estimates

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Background

Genetic connectedness quantifies the extent to which estimated breeding values can be fairly compared across units or contemporary groups [1, 2]. Genetic evaluation is known to be more robust when the connectedness level is high enough due to sufficient sharing of genetic material across groups. In such scenarios, the best linear unbiased prediction minimizes the risk of uncertainty in ranking of individuals. On the other hand, limited or no sharing of genetic material leads to less reliable comparisons of genetic evaluation methods [3]. High-throughput genetic variants spanning the entire genome available for a wide range of agricultural species have now opened up an opportunity to assess connectedness using genomic data. A recent study showed that genomic relatedness strengthens the measures of connectedness across units compared with the use of pedigree relationships [4]. The concept of genetic connectedness was later extended to measure the connectedness level between reference and validation sets in whole-genome prediction [5]. This approach has also been used to optimize individuals constituting reference sets [6, 7] In general, it was observed that increased connectedness led to increased prediction accuracy of genetic values evaluated by cross-validation [8]. Comparability of total genetic values across units by accounting for additive as well as non-additive genetic effects has also been investigated [9].

Despite the importance of connectedness, there is no user-friendly software tool available that offers computation of a comprehensive list of connectedness statistics using pedigree and genomic data. Therefore, we developed a genetic connectedness analysis R package, GCA, which measures the connectedness between individuals across units using pedigree or genomic data. The objective of this article is to describe a large collection of connectedness statistics implemented in the GCA package, overview the software architecture, and present several examples using simulated data.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Connectedness statistics

A list of connectedness statistics supported by the GCA R package is shown in Figure 1. These statistics can be classified into core functions derived from either prediction error variance (PEV) or variance of unit effect estimates (VE). PEV-derived metrics include prediction error variance of differences (PEVD), coefficient of determination (CD), and prediction error correlation (r). Further, each metric based on PEV can be summarized as the average PEV within and across units, at the unit level as the average PEV of all pairwise differences between individuals across units, or using a contrast vector. VE-derived metrics include variance of differences in unit effects (VED), coefficient of determination of VED (CDVED), and connectedness rating (CR). For each VE-derived metric, three correction factors accounting for the number of fixed effects can be applied. These include no correction (0), correcting for one fixed effect (1), and correcting for two or more fixed effects (2). Thus, a combination of core functions, metrics, summary functions, and correction factors uniquely characterizes connectedness statistics. Further, the overall connectedness statistic can be obtained by calculating the average of the pairwise connectedness statistics across units.

Core functions

Prediction error variance (PEV)

A PEV matrix is obtained from Henderson's mixed model equations (MME) by assuming a standard linear mixed model $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\epsilon}$, where $\mathbf{y}$, $\mathbf{b}$, $\mathbf{u}$, and $\boldsymbol{\epsilon}$ refer to a vector of phenotypes, systematic effects, random additive genetic effects, and residuals, respectively [10]. The $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices associating systematic effects and genetic values to observations, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Then the PEV of $\mathbf{u}$ is derived as shown in Henderson [10].

$$\begin{aligned}
\mathrm{PEV}(\mathbf{u}) &= \mathrm{Var}(\hat{\mathbf{u}} - \mathbf{u}) \\
&= \mathrm{Var}(\mathbf{u}|\hat{\mathbf{u}}) \\
&= (\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{K}^{-1}\lambda)^{-1}\sigma_\epsilon^2 \\
&= \mathbf{C}^{22}\sigma_\epsilon^2,
\end{aligned}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is the absorption (projection) matrix for fixed effects, $\mathbf{K}$ is a relationship matrix, and $\lambda = \frac{\sigma_\epsilon^2}{\sigma_u^2}$ is the ratio of residual and additive genetic variance. $\mathbf{C}^{22}$ represents the lower right quadrant of the inverse of coefficient matrix. Note that $\mathrm{PEV}(\mathbf{u})$ can be viewed as the posterior variance of $\mathbf{u}$.

Variance of unit effect estimates (VE)

An alternative option for the choice of core function is to use VE, which is based on the variance-covariance matrix of estimated unit or contemporary group effects. Kennedy and Trus [11] argued that mean PEV over unit ($\mathrm{PEV}_{\mathrm{Mean}}$) defined as the average of PEV between individuals within the same unit can be approximated by $\mathrm{VE} = \mathrm{Var}(\hat{\mathbf{b}})$, that is

$$\begin{aligned}
\mathrm{VE}_0 &= \mathrm{Var}(\hat{\mathbf{b}}) \\
&= [\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \mathbf{K}^{-1}\lambda)^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\sigma_\epsilon^2 \\
&\approx \mathrm{PEV}_{\mathrm{Mean}}
\end{aligned} \tag{1}$$

Holmes et al. [12] pointed out that the agreement between $\mathrm{PEV}_{\mathrm{Mean}}$ and VE_0 depends on a number of fixed effects other than the management group fitted in the model. They proposed exact ways to derive $\mathrm{PEV}_{\mathrm{Mean}}$ as a function of VE and suggested addition of a few correction factors. When unit

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

effect is the only fixed effect included in the model, the exact $\mathrm{PEV}_{\mathrm{Mean}}$ can be obtained as given below.

$$\mathrm{VE}_1 = \mathrm{PEV}_{\mathrm{Mean}} = \mathrm{Var}(\hat{\mathbf{b}}) - \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}, \tag{2}$$

where $\mathbf{X}'\mathbf{X}^{-1}$ is a diagonal matrix with ith diagonal element equal to $\frac{1}{n_i}$, and n_i is the number of records in unit i. Thus, the term $\sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}$ corrects the number of records within units. Accounting for additional fixed effects beyond unit effect when computing $\mathrm{PEV}_{\mathrm{Mean}}$ is given by the following equation.

$$\begin{aligned}
\mathrm{VE}_2 &= \mathrm{PEV}_{\mathrm{Mean}} && (3)\\
&= \mathrm{Var}(\hat{\mathbf{b}}_1) - \sigma_\epsilon^2(\mathbf{X_1}'\mathbf{X_1})^{-1} \\
&+ (\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{X_2}\mathrm{Var}(\hat{\mathbf{b}}_2)\mathbf{X_2}'\mathbf{X_1}(\mathbf{X_1}'\mathbf{X_1})^{-1} \\
&+ (\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{X_2}\mathrm{Cov}(\hat{\mathbf{b}}_2, \hat{\mathbf{b}}_1) \\
&+ \mathrm{Cov}(\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2)\mathbf{X_2}'\mathbf{X_1}(\mathbf{X_1}'\mathbf{X_1})^{-1}, && (4)
\end{aligned}$$

where $\mathbf{X_1}$ and $\mathbf{X_2}$ represent incidence matrices for units and other fixed effects, respectively, and $\hat{\mathbf{b}}_1$ and $\hat{\mathbf{b}}_2$ refer to the estimates of unit effects and other fixed effects, respectively [12]. This equation is suitable for cases in which there are two or more fixed effects fitted in the model.

Connectedness metrics

Below we describe connectedness metrics implemented in the GCA package. We also summarized and organized their relationships with each other, which were never clearly articulated in the literature. These metrics are the function of PEV or VE described earlier (Figure 1).

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Prediction error variance of difference (PEVD)

A PEVD metric measures the prediction error variance difference of breeding values between individuals from different units [11]. The PEVD between two individuals i and j is expressed as shown below.

$$\begin{aligned}\mathrm{PEVD}(\hat{u}_i - \hat{u}_j) &= [\mathrm{PEV}(\hat{u}_i) + \mathrm{PEV}(\hat{u}_j) - 2\mathrm{PEC}(\hat{u}_i, \hat{u}_j)] \\ &= (\mathbf{C}^{22}_{ii} - \mathbf{C}^{22}_{ij} - \mathbf{C}^{22}_{ji} + \mathbf{C}^{22}_{jj})\sigma^2_\epsilon \\ &= (\mathbf{C}^{22}_{ii} + \mathbf{C}^{22}_{jj} - 2\mathbf{C}^{22}_{ij})\sigma^2_\epsilon, \end{aligned} \tag{5}$$

where PEC_{ij} is the off-diagonal element of the PEV matrix corresponding to the prediction error covariance between errors of genetic values.

Group average PEVD: The average PEVD derived from the average relationships between and within units as a choice of connectedness measure can be traced back to Kennedy and Trus [11]. This can be calculated by inserting the $\mathrm{PEV}_{\mathrm{Mean}}$ of i'th and j'th units and mean prediction error covariance ($\mathrm{PEC}_{\mathrm{Mean}}$) between i'th and j'th units into equation (5) as

$$\mathrm{PEVD}_{i'j'} = \overline{\mathrm{PEV}}_{i'i'} + \overline{\mathrm{PEV}}_{j'j'} - 2\overline{\mathrm{PEC}}_{i'j'}, \tag{6}$$

where $\overline{\mathrm{PEV}}_{i'i'}$, $\overline{\mathrm{PEV}}_{j'j'}$, and $\overline{\mathrm{PEC}}_{i'j'}$ denote $\mathrm{PEV}_{\mathrm{Mean}}$ in i'th and j'th units, and $\mathrm{PEC}_{\mathrm{Mean}}$ between i'th and j'th units. We refer to this summary method as group average as illustrated in Figure 2A.

Individual average PEVD: Alternatively, we can first compute PEVD at the individual level using equation (5) and then aggregate and summarize at the unit level to obtain the average of PEVD between individuals across two units [13]

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$$\text{PEVD}_{i'j'} = \frac{1}{n_{i'} \cdot n_{j'}} \sum \text{PEVD}_{i'j'},$$

where $n_{i'}$ and $n_{j'}$ are the total number of records in units i' and j', respectively and $\sum \text{PEVD}_{i'j'}$ is the sum of all pairwise differences between the two units. We refer to this summary method as individual average. A flow diagram illustrating the computational procedure is shown in Figure 2B.

Contrast PEVD: The PEVD of contrast between a pair of units can be used to summarize PEVD [14].

$$\text{PEVD}(\mathbf{x}) = \mathbf{x}'\mathbf{C}^{22}\mathbf{x}\sigma_\epsilon^2,$$

where $\mathbf{x}$ is a contrast vector involving $1/n_{i'}$, $1/n_{j'}$ and 0 corresponding to individuals belonging to i'th, j'th, and the remaining units. The sum of elements in $\mathbf{x}$ equals to zero. A flow diagram showing a computational procedure is shown in Figure 2C.

Coefficient of determination (CD)

A CD metric measures the precision of genetic values and can be interpreted as the square of the correlation between the predicted and the true difference in the genetic values or the ratio of posterior and prior variances of genetic values $\mathbf{u}$ [15]. A notable difference between CD and PEVD is that CD penalizes connectedness measurements when across units include individuals that are genetically too

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

similar [4, 8]. A pairwise CD between individuals i and j is given by the following equation.

$$\begin{aligned}
\mathrm{CD}_{ij} &= \frac{\mathrm{Var}(\hat{\mathbf{u}})}{\mathrm{Var}(\mathbf{u})} \\
&= \frac{\mathrm{Var}(\mathbf{u}) - \mathrm{Var}(\mathbf{u}|\hat{\mathbf{u}})}{\mathrm{Var}(\mathbf{u})} \\
&= 1 - \frac{\mathrm{Var}(\mathbf{u}|\hat{\mathbf{u}})}{\mathrm{Var}(\mathbf{u})} \\
&= 1 - \lambda \frac{\mathbf{C}^{22}_{ii} + \mathbf{C}^{22}_{jj} - 2\mathbf{C}^{22}_{ij}}{\mathbf{K}_{ii} + \mathbf{K}_{jj} - 2\mathbf{K}_{ij}},
\end{aligned}$$

where $\mathbf{K}_{ii}$ and $\mathbf{K}_{jj}$ are ith and jth diagonal elements of $\mathbf{K}$, and $\mathbf{K}_{ij}$ is the relationship between ith and jth individuals [14].

Group average CD: Similar to the group average PEVD statistic, $\mathrm{PEV}_{\mathrm{Mean}}$ and $\mathrm{PEC}_{\mathrm{Mean}}$ can be used to summarize CD at the unit level.

$$\begin{aligned}
\mathrm{CD}_{i'j'} &= 1 - \lambda \cdot \frac{\overline{\mathbf{C^{22}}}_{i'i'} + \overline{\mathbf{C^{22}}}_{j'j'} - 2\overline{\mathbf{C^{22}}}_{i'j'}}{(\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})} \\
&= 1 - \frac{\sigma_e^2 \cdot \overline{\mathbf{C^{22}}}_{i'i'} + \overline{\mathbf{C^{22}}}_{j'j'} - 2\overline{\mathbf{C^{22}}}_{i'j'}}{\sigma_u^2 \cdot (\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})} \\
&= 1 - \frac{\overline{\mathrm{PEV}}_{i'i'} + \overline{\mathrm{PEV}}_{j'j'} - 2\overline{\mathrm{PEC}}_{i'j'}}{\sigma_u^2 \cdot (\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})} \\
&= 1 - \frac{\mathrm{PEVD}_{i'j'}}{\sigma_u^2 \cdot (\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})}. \qquad (7)
\end{aligned}$$

Here, $\overline{\mathbf{K}}_{i'i'}$, $\overline{\mathbf{K}}_{j'j'}$ and $\overline{\mathbf{K}}_{i'j'}$ refer to the means of relationship coefficients in units i' and j', and the mean relationship coefficient between two units i' and j', respectively. Graphical derivation of group average CD is illustrated in Figure 3A. This summary method has not been used in the literature, but shares the same spirit with the group average PEVD.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Individual average CD: Individual average CD is derived from the average of CD between individuals across two units [13].

$$
\begin{aligned}
\mathrm{CD}_{i'j'} &= 1 - \lambda \cdot \frac{\frac{1}{n_{i'} \cdot n_{j'}} \cdot \sum \left(\mathbf{C^{22}}_{i'i'} + \mathbf{C^{22}}_{j'j'} - 2\mathbf{C^{22}}_{i'j'} \right)}{\frac{1}{n_{i'} \cdot n_{j'}} \cdot \sum (\mathbf{K}_{i'i'} + \mathbf{K}_{j'j'} - 2\mathbf{K}_{i'j'})} \\
&= 1 - \frac{\frac{1}{n_{i'} \cdot n_{j'}} \cdot \sigma_e^2 \cdot \sum \left(\mathbf{C^{22}}_{i'i'} + \mathbf{C^{22}}_{j'j'} - 2\mathbf{C^{22}}_{i'j'} \right)}{\frac{1}{n_{i'} \cdot n_{j'}} \cdot \sigma_u^2 \cdot \sum (\mathbf{K}_{i'i'} + \mathbf{K}_{j'j'} - 2\mathbf{K}_{i'j'})} \\
&= 1 - \frac{\frac{1}{n_{i'} \cdot n_{j'}} \sum \mathrm{PEVD}_{i'j'}}{\frac{1}{n_{i'} \cdot n_{j'}} \cdot \sigma_u^2 \cdot \sum (\mathbf{K}_{i'i'} + \mathbf{K}_{j'j'} - 2\mathbf{K}_{i'j'})} \\
&= 1 - \frac{\sum \mathrm{PEVD}_{i'j'}}{\sigma_u^2 \cdot \sum (\mathbf{K}_{i'i'} + \mathbf{K}_{j'j'} - 2\mathbf{K}_{i'j'})}.
\end{aligned}
$$

A flow diagram of individual average CD is shown in Figure 3B.

Contrast CD: A contrast of CD between any pair of units is given by [14]

$$
\begin{aligned}
\mathrm{CD}(\mathbf{x}) &= 1 - \frac{\mathrm{Var}(\mathbf{x}'\mathbf{u}|\hat{\mathbf{u}})}{\mathrm{Var}(\mathbf{x}'\mathbf{u})} \\
&= 1 - \lambda \cdot \frac{\mathbf{x}'\mathbf{C}^{22}\mathbf{x}}{\mathbf{x}'\mathbf{K}\mathbf{x}} \\
&= 1 - \frac{\mathbf{x}'\mathbf{C}^{22}\mathbf{x} \cdot \sigma_e^2}{\mathbf{x}'\mathbf{K}\mathbf{x} \cdot \sigma_u^2} \\
&= 1 - \frac{\mathrm{PEVD(x)}}{\mathbf{x}'\mathbf{K}\mathbf{x} \cdot \sigma_u^2}.
\end{aligned}
$$

A flow diagram showing the computational procedure is shown in Figure 3C.

Prediction error correlation (r)

Prediction error correlation, known as pairwise r statistic, between individuals i and j is calculated from the elements of the PEV matrix [16].

$$
\mathrm{r}_{ij} = \frac{\mathrm{PEC}(\hat{u}_i, \hat{u}_j)}{\sqrt{\mathrm{PEV}(\hat{u}_i) \cdot \mathrm{PEV}(\hat{u}_j)}}.
$$

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Group average r: This is known as flock connectedness in the literature, which calculates the ratio of $\mathrm{PEV}_{\mathrm{Mean}}$ and $\mathrm{PEC}_{\mathrm{Mean}}$ [16]. This group average connectedness for r between two units i' and j' is given by

$$\begin{aligned}
\mathrm{r}_{i'j'} &= \frac{\overline{\mathrm{PEC}}_{i'j'}}{\sqrt{\overline{\mathrm{PEV}}_{i'i'} \cdot \overline{\mathrm{PEV}}_{j'j'}}} \\
&= \frac{1/n_{i'} \sum \mathrm{PEC}_{i'j'} 1/n_{j'}}{\sqrt{(1/n_{i'})^2 \sum \mathrm{PEV}_{i'i'} \cdot (1/n_{j'})^2 \sum \mathrm{PEV}_{j'j'}}} \\
&= \frac{\sum \mathrm{PEC}_{i'j'}}{\sqrt{\sum \mathrm{PEV}_{i'i'} \cdot \sum \mathrm{PEV}_{j'j'}}}.
\end{aligned} \tag{8}$$

A graphical derivation is presented in Figure 4A.

Individual average r: The summary method based on individual average calculates pairwise r for all pairs of individuals followed by averaging all r measures across units.

$$\mathrm{r}_{i'j'} = \frac{1}{n_{i'} \cdot n_{j'}} \cdot \sum \frac{\mathrm{PEC}(\hat{u}_{i'}, \hat{u}_{j'})}{\sqrt{\mathrm{PEV}(\hat{u}_{i'}) \cdot \mathrm{PEV}(\hat{u}_{j'})}}.$$

This summary method was first used in Yu et al. [4] and calculation steps are shown in Figure 4B.

Contrast r: A contrast of r is defined as below.

$$\mathrm{r}(\mathbf{x}) = \mathbf{x}'\mathbf{r}\mathbf{x}.$$

This summary method has not been used in the literature, but shares the same concept with the contrasts PEVD and CD. A flow diagram illustrating a computational procedure is shown in Figure 4C.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Variance of differences in unit effects (VED)

A metric VED, which is a function of VE can be used to measure connectedness. All PEV-based metrics follow a two-step procedure in the sense that they first compute the PEV matrix at the individual level and then apply one of the summary methods to derive connectedness at the unit level or vice versa. In contrast, VE-based metrics follow a single-step procedure such that we can obtain connectedness between units directly. Moreover, since the number of fixed effects is oftentimes smaller than the number of individuals in the model, the computational requirements for VED are expected to be lower [12]. Note that all VE-derived approaches can be classified based on the number of fixed effects to be corrected. Using the group average summary method, three VEDc statistics estimate PEVD alike connectedness between two units i' and j' by replacing $\mathrm{PEV}_{\mathrm{Mean}}$ in equation (6) from VEc [11, 12].

$$\mathrm{VEDc}_{i'j'} = \mathrm{VEc}_{i'i'} + \mathrm{VEc}_{j'j'} - 2\mathrm{VEc}_{i'j'}, \tag{9}$$

Here, c denotes no correction (0), correction for one fixed effect (1), and correction for two or more fixed effects (2) [12].

Coefficient of determination of VED

Similarly, the correction function based on VEc can be employed to define a group average CD alike statistic. We named this as CDVEDc. A pairwise CDVEDc between two units i' and j' is given by

$$\mathrm{CDVEDc}_{i'j'} = 1 - \frac{\mathrm{VEc}_{i'i'} + \mathrm{VEc}_{j'j'} - 2\mathrm{VEc}_{i'j'}}{\sigma_u^2 \cdot (\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})}.$$

Here, c includes 0, 1, and 2 by referring to the number of corrections for fixed effects.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Connectedness rating (CR)

A CR statistic first proposed by Mathur et al. [17] is similar to equation (8). However, it uses variances and covariances of estimated unit effects instead of $\mathrm{PEV}_{\mathrm{Mean}}$ and $\mathrm{PEC}_{\mathrm{Mean}}$. Holmes et al. [12] extended CR by replacing VE with VEc to calculate CR and this is referred as CRc below. A pairwise CRc between two units i' and j' is outlined as

$$\mathrm{CRc}_{i'j'} = \frac{\mathrm{VEc}_{i'j'}}{\sqrt{\mathrm{VEc}_{i'i'} \cdot \mathrm{VEc}_{j'j'}}},$$

where c equals to the number of corrections for fixed effects: 0, 1, and 2. When c is set to 0, this is equivalent to CR of Mathur et al. [17].

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Software Description

Overview of software architecture

The GCA R package is implemented entirely in R, which is an open source programming language and environment for performing statistical computing [18]. The package is hosted on a GitHub page accompanied by a detailed vignette document. Computational speed was improved by integrating C++ code into R code using the Rcpp package [19]. The initial versions of the algorithms and the R code were used in previous studies [4, 8, 9] and were enhanced further for efficiency, usability, and documentation in the current version to facilitate connectedness analysis. The GCA R package provides a comprehensive and effective tool for genetic connectedness analysis and whole-genome prediction, which further contributes to the genetic evaluation and prediction.

Installing the GCA package

The current version of the GCA R package is available at GitHub (`https://github.com/QGresources/GCA`). The package can be installed using the devtools R package [20] and loaded into the R environment following the steps shown at GitHub.

Simulated data

A simulated cattle data set using QMSim software [21] is included in the GCA package as an example data set. A total of 2,500 cattle spanning five generations were simulated with pedigree and genomic information available for all individuals. We simulated 10,000 evenly distributed biallelic single nucleotide polymorphisms and 2,000 randomly distributed quantitative trait loci (QTL) across 29 pairs of autosomes with 100 cM per chromosome. A single phenotype with a heritability of 0.6 and a fixed covariate of sex were simulated. This was followed by simulating units using the k-medoid algorithm [22] coupled with the dissimilarity matrix derived from a numerator relationship matrix as shown in

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

previous studies [4, 8, 9]. The data set is stored as an R object in the package. The genotype object is a $2,500 \times 10,000$ marker matrix. The phenotype object is a $2,500 \times 6$ matrix, including the columns of progeny, sire, dam, sex, unit, and phenotype.

Application of the GCA package

A detailed usage of the GCA R package can be found in the vignette document (https://qgresources.github.io/GCA_Vignette/GCA.html). Examples include 1) pairwise and overall connectedness measures across units; 2) relationship between PEV- and VE-based connectedness metrics; and 3) relationship between connectedness metrics and genomic prediction accuracies.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Conclusions

The GCA R package provides users with a comprehensive tool for analysis of genetic connectedness using pedigree and genomic data. The users can easily assess the connectedness of their data and be mindful of the uncertainty associated with comparing genetic values of individuals involving different management units or contemporary groups. Moreover, the GCA package can be used to measure the level of connectedness between training and testing sets in the whole-genome prediction paradigm. This parameter can be used as a criterion for optimizing the training data set. This paper also summarized the relationship among various connectedness metrics, which was not clearly articulated in the past literature. In summary, we contend that the availability of the GCA package to calculate connectedness allows breeders and geneticists to make better decisions on comparing individuals in genetic evaluations and inferring linkage between any pair of individual groups in genomic prediction.

Availability and implementation

The GCA R source code is provided as free and open source. The webpage https://github.com/QGresources/GCA was created as a nexus of all genetic connectedness related functions and examples available in the GCA R package. The vignette is available at https://qgresources.github.io/GCA_Vignette/GCA.html.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Declarations

Funding

This work was supported in part by Virginia Polytechnic Institute and State University startup funds to GM.

Authors' contributions

HY and GM developed the software tool and wrote the manuscript. GM supervised and directed the study. All authors read and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figures

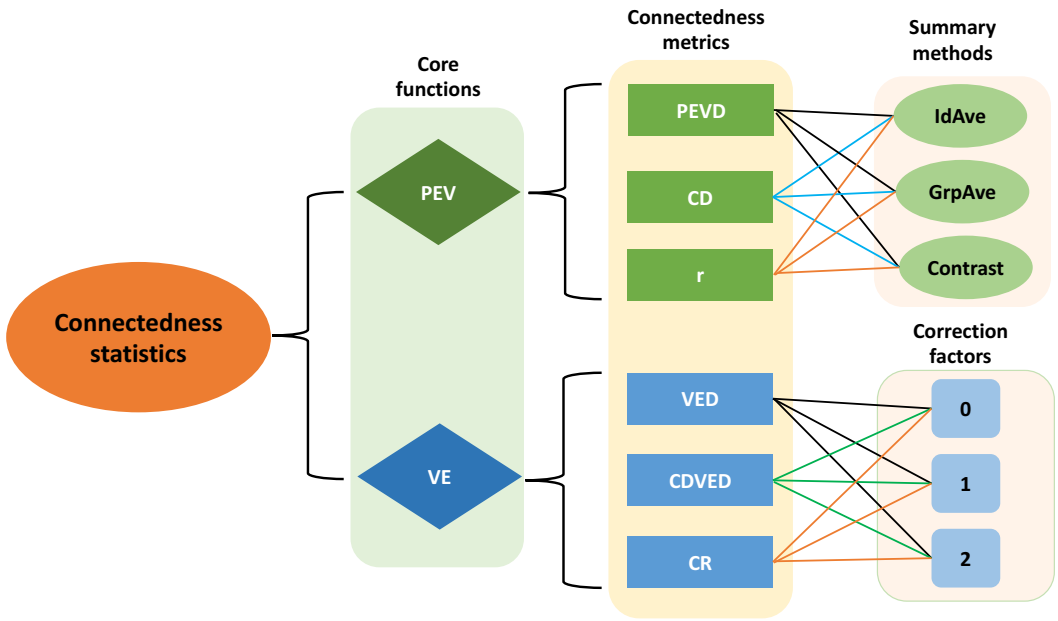

Figure 1: An overview of connectedness statistics implmented in the GCA R package. The statistics can be computed from either prediction error variance (PEV) or variance of unit effect estimates (VE). Connectedness metrics include prediction error variance of the difference (PEVD), coefficient of determination (CD), prediction error correlation (r), variance of differences in unit effects (VED), coefficient of determination of VE (CDVE), and connectedness rating (CR). IdAve, GrpAve, and Contrast correspond to individual average, group average, and contrast summary methods, respectively. 0, 1, and 2 are correction factors accounting for the fixed effects in the model.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

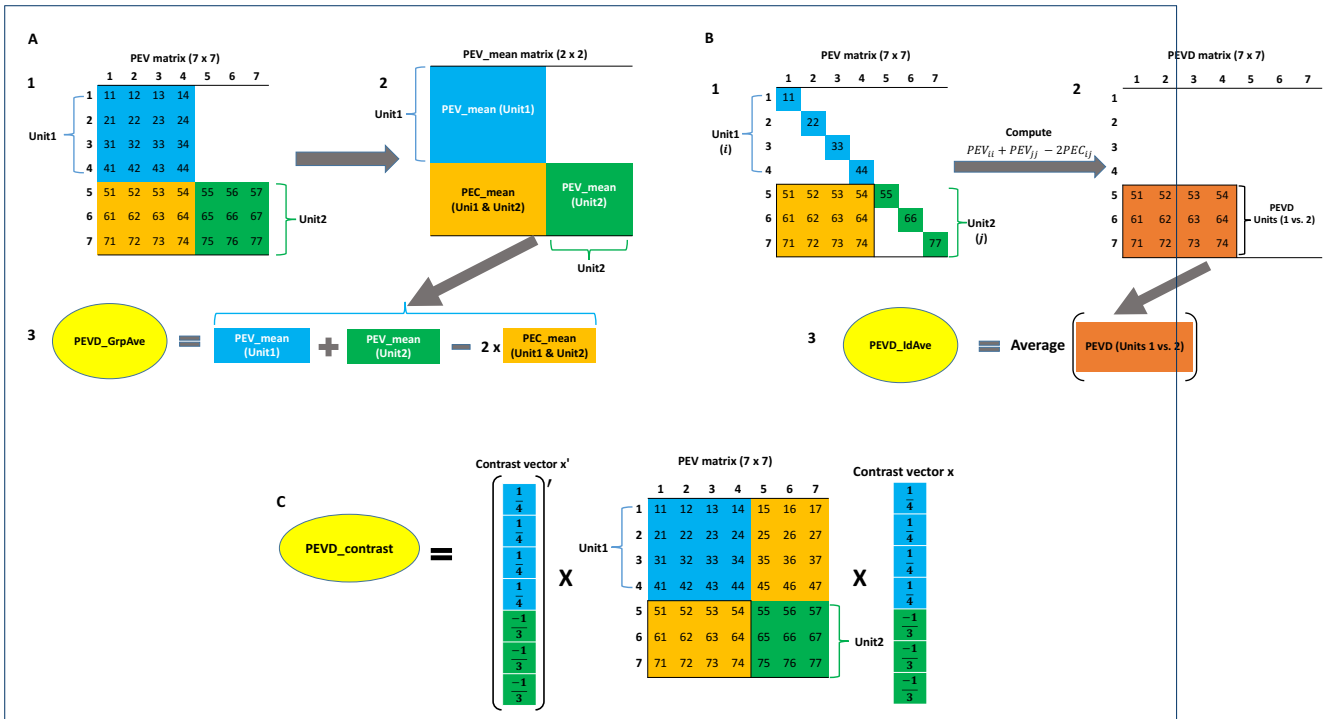

Figure 2: A flow diagram of three prediction error variance of the difference (PEVD) statistics. The group average PEVD (PEVD_GrpAve) is shown in A. A1: Prediction error variance (PEV) matrix including variances and covariances of seven individuals. A2: Calculate the mean of prediction error variance / covariance within the unit (PEV_mean) and mean of prediction error covariance across the unit (PEC_mean). A3: Group average PEVD is calculated by applying the PEVD equation using PEV_mean and PEC_mean. The individual average PEVD (PEVD_IdAve) is shown in B. B1: Prediction error variance (PEV) matrix including variances and covariances of seven individuals. Subscripts i and j refer to the ith and jth individuals in units 1 and 2, respectively. B2: Pairwise PEVD between individuals across two units. B3: Individual average PEVD is calculated by taking the average of all pairwise PEVD. The PEVD of contrast (PEVD_Contrast) is shown in C. PEVD_Contrast is calculated as the product of the transpose of the contrast vector ($\mathbf{x}$), the PEV matrix, and the contrast vector.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

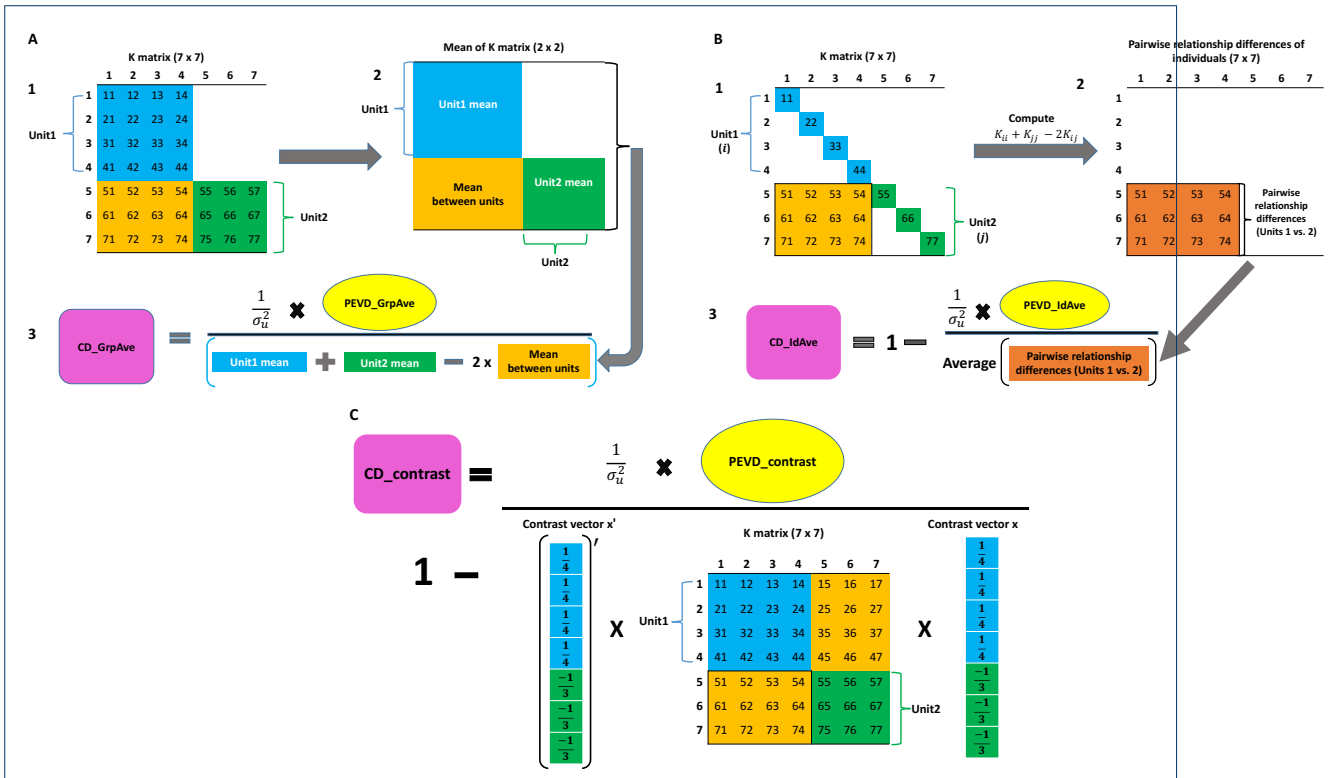

Figure 3: A flow diagram of three coefficient of determination (CD) statistics. The group average CD (CD_GrpAve) is shown in A. A1: A relationship matrix of seven individuals. A2: Calculate the mean relationships within and between units. A3: Group average CD is calculated by scaling group average PEVD (PEVD_GrpAve) by the quantity obtained from the PEVD equation using the within and between unit means. The individual average CD (CD_IdAve) is shown in B. B1: A relationship matrix of seven individuals. B2: Calculate pairwise relationship differences of individuals between the units. Subscripts i and j refer to the ith and jth individuals in units 1 and 2, respectively. B3: Individual average CD is calculated by scaling indvidual average PEVD (PEVD_IdAve) with the average of pairwise relationship differences of individuals. The CD of contrast (CD_Contrast) is shown in C. CD_Contrast is calculated by scaling the prediction error variance of the differences (PEVD) of contrast with the product of the transpose of the contrast vector ($\mathbf{x}$), the relationship matrix ($\mathbf{K}$), and the contrast vector.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

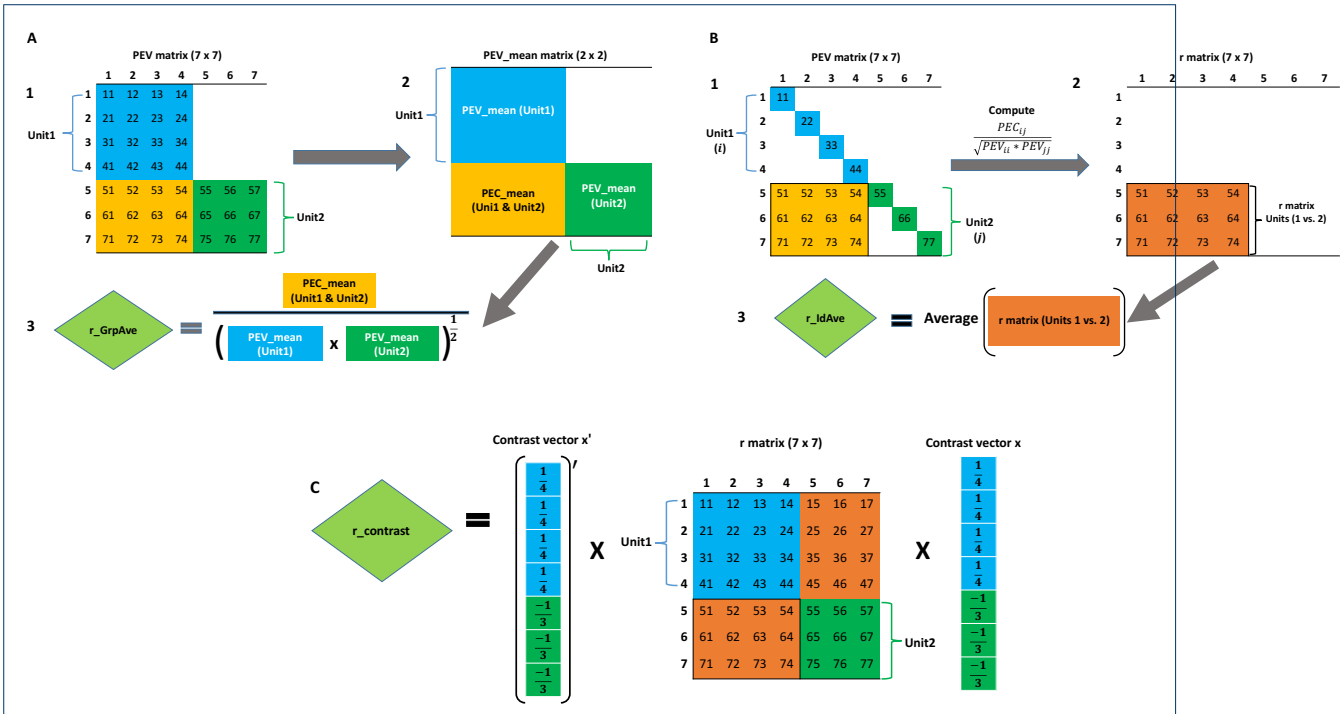

Figure 4: A flow diagram of three prediction error correlation (r) statistics. The group average r (r_GrpAve) is shown in A. A1: Prediction error variance (PEV) matrix of seven individuals. A2: Calculate the mean of prediction error variance / covariance within the unit (PEV_mean) and mean of prediction error covariance across the unit (PEC_mean). A3: Group average r is a correlation calculated from PEV_mean and PEC_mean. The calculation of individual average r (r_IdAve) involving seven individuals is displayed in B. B1: Prediction error variance (PEV) matrix of seven individuals. B2: Calculate pairwise correlation coefficients of individuals between units using PEV and prediction error covariance (PEC). Subscripts i and j refer to the ith and jth individuals in units 1 and 2, respectively. B3: Individual average r is calculated as the average of pairwise prediction error correlation coefficients of individuals across units. The r of contrast (r_Contrast) is shown in C. r_Contrast is calculated from the product of the transpose of the contrast vector ($\mathbf{x}$), r matrix, and the contrast vector.

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Author details

References

1. Foulley, J., Schaeffer, L., Song, H., Wilton, J.: Progeny group size in an organized progeny test program of ai beef bulls using reference sires. Canadian Journal of Animal Science **63**(1), 17–26 (1983)
2. Foulley, J., Bouix, J., Goffinet, B., et al.: Connectedness in genetic evaluation, pp. 277–308. Springer (1990)
3. Kuehn, L., Notter, D., Nieuwhof, G., Lewis, R.: Changes in connectedness over time in alternative sheep sire referencing schemes. Journal of Animal Science **86**(3), 536–544 (2008)
4. Yu, H., Spangler, M.L., Lewis, R.M., Morota, G.: Genomic relatedness strengthens genetic connectedness across management units. G3: Genes, Genomes, Genetics **7**(10), 3543–3556 (2017)
5. Pszczola, M., Strabel, T., Van Arendonk, J., Calus, M.: The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. Journal of dairy science **95**(9), 5412–5421 (2012)
6. Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodriguez, V.M., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., *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 **192**(2), 715–728 (2012)
7. Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., Sorrells, M.E.: Training set optimization under population structure in genomic selection. Theoretical and applied genetics **128**(1), 145–158 (2015)
8. Yu, H., Spangler, M.L., Lewis, R.M., Morota, G.: Do stronger measures of genomic connectedness enhance prediction accuracies across management units? Journal of Animal Science **96**(11), 4490–4500 (2018)
9. Momen, M., Morota, G.: Quantifying genomic connectedness and prediction accuracy from additive and non-additive gene actions. Genetics Selection Evolution **50**(1), 45 (2018)
10. Henderson, C.R.: Applications of Linear Models in Animal Breeding. University of Guelph, Third edition, Edited by Schaeffer LR. Guelph (1984)
11. Kennedy, B., Trus, D.: Considerations on genetic connectedness between management units under an animal model. Journal of Animal Science **71**(9), 2341–2352 (1993)
12. Holmes, J.B., Dodds, K.G., Lee, M.A.: Estimation of genetic connectedness diagnostics based on prediction errors without the prediction error variance–covariance matrix. Genetics Selection Evolution **49**(1), 29 (2017)
13. Amorim, S.T., Yu, H., Baldi, F., Morota, G.: An assessment of genomic connectedness measures in nellore cattle. implementation and strategies for national beef cattle genetic evaluation. In: NCERA225: Implementation and Strategies for National Beef Cattle Genetic Evaluation (2019)
14. Laloë, D., Phocas, F., Menissier, F.: Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. Genetics Selection Evolution **28**(4), 359 (1996)
15. Laloë, D.: Precision and information in linear models of genetic evaluation. Genetics Selection Evolution **25**(6), 557 (1993)
16. Lewis, R., Crump, R., Simm, G., Thompson, R.: Assessing connectedness in across-flock genetic evaluations. Proc. Brit. Soc. Anim. Sci **121** (1999)
17. Mathur, P., Sullivan, B., Chesnais, J.: Measuring connectedness: concept and application to a large industry breeding program. In: Proc. 7th World Congr. Genet. Appl. to Livest. Prod, vol. 19, p. 23 (2002)
18. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019). R Foundation for Statistical Computing. https://www.R-project.org/
19. Eddelbuettel, D., François, R.: Rcpp: Seamless R and C++ integration. Journal of Statistical Software **40**(8), 1–18 (2011). doi:10.18637/jss.v040.i08
20. Wickham, H., Chang, W.: Devtools: Tools to make developing r packages easier. R package version **1**(0), 9000 (2016)

bioRxiv preprint doi: https://doi.org/10.1101/696419; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

21. Sargolzaei, M., Schenkel, F.S.: Qmsim: a large-scale genome simulator for livestock. Bioinformatics **25**(5), 680–681 (2009)

22. Kaufman, P. L & Rousseeuw: Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley and Sons, New York (1990)