Designing and modeling high-throughput phenotyping data in quantitative genetics

Haipeng Yu

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Animal and Poultry Sciences

Gota Morota, Chair
Heather Bradford
Ina Hoeschele
M. A. Saghai Maroof
David R. Notter

March 18th, 2020
Blacksburg, Virginia

Keywords: bayesian network, factor analysis, genomic connectedness, genomic prediction, high-throughput phenotyping data

Designing and modeling high-throughput phenotyping data in quantitative genetics

Haipeng Yu

(ABSTRACT)

Quantitative genetics aims to bridge the genome to phenome gap. The advent of high-throughput genotyping technologies has accelerated the progress of genome to phenome mapping, but a challenge remains in phenotyping. Various high-throughput phenotyping (HTP) platforms have been developed recently to obtain economically important phenotypes in an automated fashion with less human labor and reduced costs. However, the effective way of designing HTP has not been investigated thoroughly. In addition, high-dimensional HTP data bring up a big challenge for statistical analysis by increasing computational demands. A new strategy for modeling high-dimensional HTP data and elucidating the interrelationships among these phenotypes are needed. Previous studies used pedigree-based connectetdness statistics to study the design of phenotyping. The availability of genetic markers provides a new opportunity to evaluate connectedness based on genomic data, which can serve as a means to design HTP. This dissertation first discusses the utility of connectedness spanning in three studies. In the first study, I introduced genomic connectedness and compared it with traditional pedigree-based connectedness. The relationship between genomic connectedness and prediction accuracy based on cross-validation was investigated in the second study. The third study introduced a user-friendly connectedness R package, which provides a suite of functions to evaluate the extent of connectedness. In the last study, I proposed a new statistical approach to model high-dimensional HTP data by leveraging the combination of confirmatory factor analysis and Bayesian network. Collectively, the results from the first

three studies suggested the potential usefulness of applying genomic connectedness to design HTP. The statistical approach I introduced in the last study provides a new avenue to model high-dimensional HTP data holistically to further help us understand the interrelationships among phenotypes derived from HTP.

Designing and modeling high-throughput phenotyping data in quantitative genetics

Haipeng Yu

(GENERAL AUDIENCE ABSTRACT)

Quantitative genetics aims to bridge the genome to phenome gap. With the advent of genotyping technologies, the genomic information of individuals can be included in a quantitative genetic model. A new challenge is to obtain sufficient and accurate phenotypes in an automated fashion with less human labor and reduced costs. The high-throughput phenotyping (HTP) technologies have emerged recently, opening a new opportunity to address this challenge. However, there is a paucity of research in phenotyping design and modeling high-dimensional HTP data. The main themes of this dissertation are 1) genomic connectedness that could potentially be used as a means to design a phenotyping experiment and 2) a novel statistical approach that aims to handle high-dimensional HTP data. In the first three studies, I first compared genomic connectedness with pedigree-based connectedness. This was followed by investigating the relationship between genomic connectedness and prediction accuracy derived from cross-validation. Additionally, I developed a connectedness R package that implements a variety of connectedness measures. The fourth study investigated a novel statistical approach by leveraging the combination of dimension reduction and graphical models to understand the interrelationships among high-dimensional HTP data.

Dedication

To my parents.

Acknowledgments

My Ph.D. journey includes two chapters. I started my Ph.D. at the University of Nebraska-Lincoln (UNL) and spent the first two years there. Then I moved to Virginia Polytechnic Institute and State University (Virginia Tech) with my advisor Gota Morota. Because I studied and worked at the two universities, I obtained more chances to meet a lot of great scientists and make many friends who enriched my professional and personal life.

To my advisor Gota Morota, who brought me into the world of quantitative genetics, I owe the deepest appreciation to him for his guidance, support, and patience. During the past four years, he has spent unlimited time and effort to mentor and guide me in both work and life. His intelligence and passion for research have always encouraged me to move forward during this journey. I always feel so fortunate to have you as my advisor and being a part of the Morota's lab. Also, I would like to thank Koeun Choi, thank you for your supports and advice on my career goal.

I would like to express my gratitude to my committee at Virginia Tech: Drs. Heather Bradford, Ina Hoeschele, David Notter, and M. A. Saghai Maroof, thank you for your valuable suggestions, teaching, and willingness to be on my committee. I also wish to thank faculty members at UNL: Drs. Ronald Lewis and Matthew Spangler, thank you for your support to my genomic connectedness studies. To Dr. Harkamal Walia, thank you for your advice and for providing the rice high-throughput phenotyping data for my study.

Also, I am grateful for my fellow graduate students at UNL and Virginia Tech. Especially,

the graduate students in the Department of Animal and Poultry Sciences in my office have enriched my graduate student life at Virginia Tech. Special thanks to Chi Zhang in China, who encouraged and supported me to pursue my career dream.

Lastly, I would like to express my gratitude to my family in China for always encouraging and supporting me. To my mom and dad, thank you for always being there, loving me, supporting me, and encouraging me to chase my dream. I feel extremely blessed to be your son.

Contents

List of Figures

List of Tables

List of Abbreviations

BCFA Bayesian Confirmatory Factor Analysis

BLUP Best Linear Unbiased Prediction

BN Bayesian Network

CD Coefficient of Determination

CDVED Coefficient of Determination of VED

CFA Confirmatory Factor Analysis

CR Connectedness Rating

CV Cross-Validation

DAG Directed Acyclic Graph

EBV Estimated Breeding Values

FA Factor Analysis

HTP High-Throughput Phenotyping

IBD Identical by Descent

IBS Identical by State

LD Linkage Disequilibrium

LiDAR Light Detection and Ranging

MAF Minor Allele Frequency

MME Mixed Model Equations

MTM Multi-Trait Model

MU Management Unit

PC Principal Component

PCA Principal Component Analysis

PEC Pediction Error Covariance

PEV Prediction Error Variance

PEV_Mean PEV Over Unit

PEVD PEV Differences

PSRF Potential Scale Reduction Factor

r Prediction Error Correlation

SNPs Single Nucleotide Polymorphisms

VE Variance of Units Effects Estimates

VED Variance of Differences in Unit Effects

Chapter 1

Introduction

1.1 Background

Quantitative genetics is the study of associating genotypes with phenotypes through genetic similarity among individuals [1]. Genomic prediction is a means to predict the genomic estimated breeding values by regressing phenotypes on whole-genome single nucleotide polymorphisms markers [2]. Because the cost of high-throughput genotyping is dramatically decreasing, genomic prediction has been widely applied to animal and plant breeding. This resulted in increased prediction accuracy and accelerated genetic gains by shortening breeding cycles [3]. However, the development and implementation of phenotyping techniques have not kept pace with the ability to generate genomic data. Because the accuracy of genomic prediction relies on both phenotypes and genotypes [4, 5], the next challenge is to establish a less labor-intensive and cost-effective phenotyping framework. Recently, various high-throughput phenotyping (HTP) technologies have been proposed to produce diverse and high-dimensional phenotypes that can be utilized in genomic analysis [6, 7, 8]. This creates a situation where the dimensions of both phenotypes and genotypes are large. However, we currently lack tools for informing a better decision for phenotyping and for handling big phenotypic data in quantitative genetic analysis. Therefore, this dissertation presents genomic connectedness by estimating the extent of connectedness measures using single nucleotide polymorphisms that can be used for evaluating a phenotyping design and introduces

a new statistical approach for incorporating high-dimensional genotyping and phenotyping data in a quantitative genetic framework.

1.2 Outline of Dissertation

The outline of this dissertation is as below. In Chapter 2, I reviewed the application of connectedness in animal and plant breeding and currently available statistical models to incorporate high-dimensional phenotypes in quantitative genetics. Chapter 3 assesses the impact of genome-wide marker information on the estimates of connectedness across units. Various kernel matrices and connectedness statistics were used to evaluate the gains in the estimates of connectedness when moving from pedigree to genome-wide markers by leveraging the combination of simulated and real data. In Chapter 4, I investigated the relationship between the measures of genomics connectedness and genome-enabled prediction accuracy based on cross-validation. I used simulated data to estimate the connectedness and genomic prediction accuracy. Despite the concept of connectedness being important in quantitative genetics, there is no user-friendly software tool available to calculate connectedness statistics. Therefore, Chapter 5 introduces the GCA R package, which utilizes pedigree or genomic data to measure the connectedness between individuals across units. This GCA R package implements a large collection of connectedness statistics derived from prediction error variance or variance of unit effect estimates. In Chapter 6, I proposed a new statistical approach to decipher genetic interrelationships among image-derived HTP data using factor analysis and graphical models. The proposed statistical framework was validated with a rice dataset to predict the potential influence on target traits when external intervention or selection was applied in the interrelated complex traits systems. The last Chapter presents the concluding remarks.

Chapter 2

Related work

2.1 Connectedness

Genetic connectedness is a measure of genetic relatedness, which assesses the comparability of estimated breeding values (EBV) from the best linear unbiased prediction across management units [9]. Genetic evaluation across units is known to be more robust when there is a sufficient level of connectedness. Otherwise, an insufficient level of connectedness will increase the risk of uncertainty in EBV comparisons because genetic signals will not be clearly separated from noise. Genetic connectedness has been traditionally studied with pedigree-based best linear unbiased prediction [10, 11]. These studies showed that the common reference sire system increases the extent of connectedness, and this allows a more accurate comparison of EBV across units [12, 13, 14]. Connectedness statistics can also be applied to assess the linkage between training and testing sets in genomic prediction. This approach has been used to optimize training sets in animals [15, 16] and plants [16, 17, 18]. These studies concluded that prediction performance could be improved when the connectedness between the training and testing sets is increased. Although the importance of connectedness in genetic evaluation has been well accepted in the literature, limited attention has been paid to study the connectedness measures using whole-genome single nucleotide polymorphisms in the context of genomic prediction.

2.1.1 Connectedness statistics

The commonly used connectedness statistics are typically derived as a function of prediction error variance (PEV), which can be obtained from Henderson's mixed model equations [19]. Foulley et al. [9] formally introduced the concept of connectedness in genetic evaluation from both theoretical and applied points of view using a sire model. Therein, a connectedness index [9] was proposed to evaluate connectedness measures. Kennedy and Trus [20] argued that the most appropriate way to measure connectedness is by using average PEV differences (PEVD) between individuals across management units. They extended the concept of connectedness from a sire model to an animal model. Laloë [21] extended the individual coefficient of determination to evaluate the precision of EBV comparisons between individuals across management units, which can be considered as the correlation between true breeding values and EBV. Lewis et al. [22] proposed the average prediction error correlation (r) to assess the connectedness across flocks.

The PEV-based connectedness statistics are all based on PEV, which is a computationally intensive task. As an alternative, Kennedy and Trus [20] suggested the use of variance of unit effects estimates (VE) to infer connectedness, where management units are treated as a random effect. They found a high correlation between the VE-based variance of differences in unit effects (VED) and PEVD. VE-based statistics are more efficient in terms of computation when the number of management units is small. However, Holmes et al. [23] showed that the estimates of PEV mean using VE is biased; therefore, VE could be a poor estimator of average PEV under certain cases. In their study, they introduced three correction functions to VE and calculated VED statistics and connectedness rating. The performance of these correction functions was evaluated by comparing with analogous PEV-based connectedness statistics, such as PEVD and r. They concluded that these correction functions can successfully adjust the fixed effects included in the model, especially when more than one fixed effects are

presented in the model.

2.2 High-throughput phenotyping

2.2.1 High-throughput phenotyping technology

Phenotyping a large number of individuals was a challenge in terms of time and labor until recently. With the improvement of digital technologies, high-throughput phenotyping (HTP) offers a non-destructive and automatic way to phenotype animals and plants. In general, HTP is useful to collect hard to measure phenotypes by traditional methods. Some commonly used image-based HTP technologies include visible light, spectral, fluorescence, thermal infrared, and three-dimensional (3D) sensor. Due to the low cost and easy use, visible light imaging has been widely used in animals and plants, such as measuring the color of tomato [24] and meat [25]. Spectral imaging combines spectroscopy and photography to produce images with many wavelength bands, including multi-spectral and hyper-spectral imaging. Spectral imaging has been used to diagnose animal disease [26] and estimate canopy moisture contents [27]. Dórea et al. [28] used milk spectra data to improve the prediction of feed intake in dairy cows. Fluorescence provides an efficient way of capturing the activities of physiological and pathological traits, such as detecting drought and salinity resistance of plants [29] and analyzing dairy products [30]. Thermal infrared imaging measures the temperature variations on the surface, such as water deficit based on plant canopy temperature [31] and detects the disease of animals [32, 33, 34]. One of the widely used 3D sensors is light detection and ranging (LiDAR), which estimates the distance between object and sensor using time of flight technology. In general, LiDAR measures the time of the laser light emitted from the sensor to the object and back to the sensor. The distance between

sensor and object reflects the shape and surface features of the object that can be used for 3D construction. For instance, LiDAR has been deployed to study the segmentation of plant organ [35, 36], detection and mapping of plants from the ground [37], and species diversity of animals [38]. A depth sensor is another widely used 3D sensor, which estimates the depth of the object by analyzing the infrared light reflected from the target object using stereo technology. McCormick et al. [39] used a 3D construction derived from the depth sensor to identify quantitative trait loci of shoot architecture in sorghum. Xia et al. [40] deployed a depth sensor to automatically detect plant leaves using 3D segmentation. In animals, Zhu et al. [41] monitored and assessed the growth of pigs using the depth camera. Fernandes et al. [42] used the depth sensor and computer vision to predict the body weight of pigs using the morphological descriptors extracted from images. Additionally, Cominotte et al. [43] predicted body weight and average daily gain of beef cattle using a depth camera.

2.2.2 Modeling high-throughput phenotyping data

Plenty of novel phenotypes are currently being produced using HTP technologies. These high-dimensional HTP data open a new opportunity to enhance the phenome to genome mapping research, but have brought substantial computational challenges for statistical genetic analysis. Moreover, understanding the interrelationships among these new types of phenotypes is crucial. Here, I review statistical approaches that have been applied to model high-dimensional data, specifically image-based HTP data. A multi-trait model (MTM) has been widely used in quantitative genetics to model correlated traits by taking advantage of the genetic or environmental covariances among phenotypes [44]. Studies that applied MTM have shown the benefit of analyzing lowly heritable or scarcely recorded traits jointly with highly heritable or densely recorded traits, provided there are genetic correlations [45]. Jia and Jannink [46] showed that marker-based Bayesian MTM improves the prediction accu-

racy compared to a univariate counterpart using both simulated and real data. Calus and Veerkamp [47] proposed three MTM by combining scarcely recorded traits with genetically correlated traits. They found that MTM improved the prediction accuracy compared to a single trait model using simulated data. Jarquín et al. [48] and Lopez-Cruz et al. [49] employed MTM to investigate gene by environmental effects in genomic prediction using high-dimensional genomic and environmental data. However, a limitation of computational efficiency arises when the number of traits is large. One common approach used to model high-dimensional data is factor analysis that assumes observed phenotypes are products of underlying unobserved factors or latent variables [50]. As the number of underlying factors is usually much less than the number of observed phenotypes, factor analysis significantly decreases computational demands. de los Campos and Gianola [50] incorporated a factor structure into multivariate analysis using one underlying common factor to represent five repeated measures of milk yield in dairy cattle. Similarly, Peñagaricano et al. [51] used confirmatory factor analysis to infer five latent variables from 19 phenotypic traits using predetermined biological classes. Further, although MTM can identify associations among traits, it does not detect the direction of these relationships. Bayesian network (BN) is a probabilistic graphical model that represents the conditional dependencies among traits using a directed acyclic graph [52]. A BN includes nodes and directed edges, which are variables and conditional dependencies among variables, respectively. A BN offers an efficient way to infer directed interrelationships among multiple variables. Morota et al. [53] used BN to investigate the pattern of Linkage Disequilibrium in Holstein cattle. Xavier et al. [54] employed the Gaussian undirected graphical model to study the interrelationships among phenotypes, genetics, and environments in soybean traits. Töpner et al. [55] used BN inferred from both genomic and residual information to analyze the multiple maize traits.

Chapter 3

Genomic Relatedness Strengthens Genetic Connectedness Across Management Units

3.1 Abstract

Genetic connectedness refers to a measure of genetic relatedness across management units (e.g., herds and flocks). With the presence of high genetic connectedness in management units, best linear unbiased prediction (BLUP) is known to provide reliable comparisons between estimated genetic values. Genetic connectedness has been studied for pedigree-based BLUP; however, relatively little attention has been paid to using genomic information to measure connectedness. In this study, we assessed genome-based connectedness across management units by applying prediction error variance of difference (PEVD), coefficient of determination (CD), and prediction error correlation (r) to a combination of computer simulation and real data (mice and cattle). We found that genomic information (**G**) increased the estimate of connectedness among individuals from different management units compared to that based on pedigree (**A**). A disconnected design benefited the most. In both datasets, PEVD and CD statistics inferred increased connectedness across units when using **G**- rather than **A**-based relatedness suggesting stronger connectedness. With r once using

allele frequencies equal to one-half or scaling **G** to values between 0 and 2, which is intrinsic to **A**, connectedness also increased with genomic information. However, PEVD occasionally increased, and r decreased when obtained using the alternative form of **G**, instead suggesting less connectedness. Such inconsistencies were not found with CD. We contend that genomic relatedness strengthens measures of genetic connectedness across units and has the potential to aid genomic evaluation of livestock species.

3.2 Introduction

The problem of connectedness or disconnectedness is particularly important in genetic evaluation of managed populations such as domesticated livestock. When selecting among animals from different management units (e.g., herds and flocks), caution is needed; choosing one animal over others across management units may be associated with greater uncertainty than selection within management units. Such uncertainty is reduced if individuals from different management units are genetically linked or connected. In such a case, best linear unbiased prediction (BLUP) offers meaningful comparison of the breeding values across management units for genetic evaluation [e.g., 56].

Structures of breeding programs have a direct influence on levels of connectedness. Wide use of artificial insemination programs generally increases genetic connectedness across management units. For example, dairy cattle populations are considered highly connected due to dissemination of genetic material from a small number of highly selected sires. The situation may be different for species with less use of artificial insemination and more use of natural service mating such as for beef cattle or sheep populations. Under these scenarios, the magnitude of connectedness across management units is reduced and genetic links are largely confined within management units.

Pedigree-based genetic connectedness has been evaluated and applied in practice [e.g., 10, 11]. However, there is a relative paucity of use of genomic information such as single nucletide polymorphisms (SNP) to ascertain connectedness. It still remains elusive in what scenarios genomics can strengthen connectedness and how much gain can be expected relative to use of pedigree information alone. Connectedness statistics have been used to optimize selective genotyping and phenotyping in simulated livestock [15] and plant populations [17], and in real maize [57, 58], and real rice data [58]. These studies concluded that the greater the connectedness between the reference and validation populations, the greater the predictive performance. However, 1) connectedness among different management units and 2) differences in connectedness measures between pedigree and genomic relatedness were not explored in those studies. For better understanding of genome-based connectedness, it is critical to examine how the presence of management units comes into play. For instance, genomic relatedness provides relationships between distant individuals that appear disconnected according to the available pedigree information. In addition, it captures Mendelian sampling that is not present in pedigree relationships [59]. Thus, genomic information is expected to strengthen measures of connectedness, which in turn refines comparisons of genetic values across different management units. The objective of this study was to assess measures of genetic connectedness across management units with use of genomic information. We leveraged the combination of real data and computer simulation to compare gains in measures of connectedness when moving from pedigree to genomic relationships. First, we studied a heterogenous mice dataset stratified by cage. Then we investigated approaches to measure connectedness using real cattle data coupled with simulated management units to have greater control over the degree of confounding between fixed management groups and genetic relationships.

3.3 Materials and Methods

3.3.1 Mice data

We analyzed a heterogeneous stock mouse population established for quantitative trait mapping [60, 61]. It was originally derived from eight inbred strains (DBA/2J, C3H/HeJ, AKR/J, A/J, BALB/cJ, CBA/J, C57BL/6J, and LP/J), followed by 50 generations of pseudorandom mating. This process introduced recombinants that allow high-resolution mapping [60, 61]. This population was used for one of the first empirical applications of genomic selection in animals [62] and later used for an array of quantitative genetic studies. The data consisted of 1,884 individuals from 169 full-sib families with approximately 11 siblings per family. Each individual was genotyped with 10,946 SNP yet none of the full-sib parents were genotyped. We removed SNP with a minor allele frequency (MAF) less than 0.05, resulting in 10,339 markers for analysis. The mice were reared in 523 cages or management units that created shared environments. The majority of full-sibs were housed in the same cages and distributed to three cages on average, i.e., a full-sib family was typically reared together in three cages. Pedigree relationships within and across full-sib individuals were 0.5 and 0, respectively. This resulted in an extreme case of genetic disconnectedness across management units. Thus, the extent of connectedness was determined by the presence or absence of full-sibs in different management units.

3.3.2 Cattle data

Pedigree information of dairy cattle was available on 1,929 cattle collected over six generations starting from a base generation 0 to generation 5 [63]. Among those, 500 individuals, mostly coming from generations 2 and 3 ($> 90\%$), had both phenotypes and genotypes. His-

toric pedigree information in addition to the 500 individuals are a source of connectedness as the pedigree-based relationship matrix was constructed from the entire pedigree. The 500 individuals were genotyped for 7,250 SNP markers. The average missing rate of genotypes across the entire SNP was 0.0002. We imputed missing genotypes by sampling alleles from a Bernoulli distribution with the marginal allele frequency used as a parameter. We retained 6,714 SNP after removing markers with MAF less than 0.05. We simulated management units in two steps: 1) individuals were clustered and 2) clusters were assigned to management units. The k-medoid clustering was performed to cluster individuals into distinctive groups. In particular, we used partitioning around medoids, which is considered a robust version of K-means [64, 65]. We formed sets of clusters so that individuals in the same groups were more similar to each other than to those in other groups. We selected the number of clusters by optimum average silhouette width algorithm implemented in the cluster and fpc R packages. This algorithm minimizes dissimilarity measures among individuals within the same cluster using the Euclidean metric and finds the optimal number of clusters that returns the lowest average dissimilarity computed from each cluster. The clustering was based on the **A** matrix, which was converted to a dissimilarity matrix by calculating the distance from the highest similarity to each similarity value in such a way that the relationship with the largest value becomes zero. We simulated the four following scenarios.

- Scenario 1: Completely disconnected - all clusters allocated to their own management units

- Scenario 2: Disconnected - one-half of clusters allocated to management unit 1 and remaining half assigned to management unit 2

- Scenario 3: Partially connected - approximately one-third of clusters allocated to management unit 1, another one-third to management unit 2, and the remaining one-third

of clusters assigned to both managements to act as a link to connect the two management units indirectly

- Scenario 4: Connected - all clusters equally allocated to the two management units

Subsequently, appropriate incidence matrices were constructed and we computed connectedness statistics across management units employing pedigree and genomic relationships.

3.3.3 Prediction error variance

Genetic connectedness statistics are typically defined as a function of the inverse of the coefficient matrix. For instance, Kennedy and Trus [20] proposed a genetic connectedness measure as the average prediction error variance (PEV) of the difference in predicted genetic values between all pairs of individuals in different management units. The PEV can be obtained from Henderson's mixed model equations (MME) [19]. We constructed MME according to a standard linear mixed model $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is a vector of phenotypes, $\mathbf{X}$ is an incidence matrix of management units, $\mathbf{b}$ is a vector of effects of management units, $\mathbf{Z}$ is an incidence matrix relating individuals to phenotypic records, $\mathbf{u}$ is a vector of random additive genetic effects, and $\boldsymbol{\epsilon}$ is a vector of residuals. The phenotypic vector $\mathbf{y}$ was standardized to have mean of 0 and variance of 1 so that results can be compared across different scenarios. The variance-covariance structure for this model is

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N \left[\begin{pmatrix} \mathbf{Xb} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{ZK}\sigma_u^2\mathbf{Z}' + \mathbf{I}\sigma_\epsilon^2 & \mathbf{ZK}\sigma_u^2 & \mathbf{I}\sigma_\epsilon^2 \\ \mathbf{KZ}'\sigma_u^2 & \mathbf{K}\sigma_u^2 & 0 \\ \mathbf{I}\sigma_\epsilon^2 & 0 & \mathbf{I}\sigma_\epsilon^2 \end{pmatrix} \right].$$

where σ_u^2 is the genetic variance, σ_e^2 is the residual variance, and $\mathbf{K}$ is a positive (semi)definite relationship matrix defined later.

The inverse of the MME coefficient matrix is represented as

$$\begin{aligned}\mathbf{C}^{-1} &= \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{K}^{-1}\lambda \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix}\end{aligned}$$

where λ is the ratio of variance components $\frac{\sigma^2_\epsilon}{\sigma^2_u}$. The PEV of genetic value for the ith individual ($\hat{u}_i$) is given by

$$\begin{aligned}\text{PEV}_i &= Var(\hat{u}_i - u_i) \\ &= Var(u_i|\hat{u}_i) \\ &= Var(\hat{u}_i|u_i) \\ &= \mathbf{C}^{22}_{ii}\sigma^2_\epsilon,\end{aligned}$$

where $\mathbf{C}^{22}_{ii}$ is the ith diagonal element of $\mathbf{C}^{22}$ coefficient matrix. Note that PEV can be interpreted as the proportion of additive genetic variance not accounted for by the prediction. Equivalently, the matrix of PEV can be computed as

$$\begin{aligned}\text{PEV} &= (\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{K}^{-1}\lambda)^{-1}\sigma^2_\epsilon \\ &= \mathbf{C}^{22}\sigma^2_\epsilon,\end{aligned}$$

where $\mathbf{M}$ is the absorption (projection) matrix for fixed effects where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$, which is orthogonal to the vector space defined by $\mathbf{X}$ (i.e., $\mathbf{M}\mathbf{X} = 0$). This avoids calculating the inverse of the entire coefficient matrix, which is useful when the number of columns of $\mathbf{X}$ is large or analysis involves repeated computation of PEV.

3.3.4 Genetic connectedness

We computed three genetic connectedness statistics: the PEV of the difference (PEVD) between genetic values [20], the coefficient of determination (CD) of the difference between predicted genetic values [21], and the prediction error correlation (r) between genetic values of individuals from different management units [66]. The first two statistics were originally used to evaluate the accuracy of individual estimated breeding values (EBV) and later extended to assess inherent risk in comparing individuals across management units. First, genetic connectedness between two individuals, i and j, was measured as PEVD [20]

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

where PEC_{ij} is the prediction error covariance or covariance between errors of genetic values, which is the off-diagonal element of the PEV matrix. If PEVD is small, individuals are said to be connected. The idea behind using PEVD as a measure of connectedness is that the accurately estimated genetic values of individuals have smaller PEV and that the pairs of genetically related individuals in the different management units have a positive prediction error covariance. Throughout this study, we used a scaled PEVD following Kuehn et al. [14] by scaling PEVD by the additive genetic variance to express connectedness without units of measurement.

Similarly, CD is closely related to PEVD and is defined by scaling the inverse of the coefficient matrix by corresponding coefficients from the relationship matrix. We can view CD as the squared correlation or reliability between the predicted and the true difference in the breeding

values [67]. This is given by

$$\mathrm{CD}_{ij} = 1 - \lambda \frac{\mathbf{C}^{22}_{ii} + \mathbf{C}^{22}_{jj} - 2\mathbf{C}^{22}_{ij}}{\mathbf{K}_{ii} + \mathbf{K}_{jj} - 2\mathbf{K}_{ij}}.$$

for pairwise comparison. In contrast to PEVD, CD accounts for the reduction of connectedness due to relationship variability between individuals under comparison. This statistic is bounded between 0 to 1, with larger values indicating increased connectedness.

The r is obtained by transforming a PEV matrix into predictive error correlation matrix. For individuals i and j, this statistic is given by

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

The rationale behind r is that there is no connectedness when PEC is zero [66]. Similar to CD, r is also bounded between 0 and 1. The larger the r, the greater the connectedness.

3.3.5 Connectedness summary

We can generalize connectedness between any pair of management units $i^{'}$ and $j^{'}$ by setting up a corresponding contrast vector $\mathbf{x}$ that sums to zero (i.e., $\mathbf{1}'\mathbf{x} = 0$) [21]. The PEVD of contrast $\mathbf{x}$ in genetic values is given by

$$\mathrm{PEVD}(\mathbf{x}) = \mathbf{x}'\mathbf{C}^{22}\mathbf{x}\sigma^2_\epsilon,$$

where $\mathbf{x}$ is a column vector including $1/n_{i'}$, $-1/n_{j'}$, and 0, for the elements corresponding to i'th unit, j'th unit, and the remaining units, respectively, where $n_{i'}$ and $n_{j'}$ were the numbers of individuals belonging to i'th and j'th units, respectively. In a contrast vector notation,

pairwise CD between management units i' and j' is given by

$$\mathrm{CD}(\mathbf{x}) = 1 - \lambda \frac{\mathbf{x}'\mathbf{C}^{22}\mathbf{x}}{\mathbf{x}'\mathbf{K}\mathbf{x}}$$

For the r statistic, a similar summary statistic can be derived as

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

where $\sum \mathrm{PEC}_{i'j'}$, $\sum \mathrm{PEV}_{i'i'}$, and $\sum \mathrm{PEV}_{j'j'}$ were the sums of the elements of $\mathrm{PEC}_{i'j'}$, $\mathrm{PEC}_{i'i'}$, and $\mathrm{PEC}_{j'j'}$, respectively [14]. However, in the Appendix A we show that when this summary statistic is applied across units it provides a reasonable summary for a pedigree relationship matrix but it is not suitable for a genomic relationship matrix when the total number of management units is two. Thus, we reported connectedness by averaging the r statistic for all pairs of individuals across management units.

3.3.6 Relationship matrix

Connectedness is realized through a genetic relationship matrix under the BLUP framework. Three genetic connectedness statistics defined above require information about covariance structures among individuals or genetic values that evaluate relatedness. We considered five $n \times n$ relationship kernel matrices ($\mathbf{K}$) in this study, where n is the number of individuals. The numerator relationship matrix, $\mathbf{K} = \mathbf{A}$, is based on relatedness due to expected additive genetic inheritance. This can be computed directly from pedigree information, and reflects the probability that alleles are inherited from a common ancestor and thereby are identical by descent (IBD). The off-diagonal elements are twice the kinship coefficients and

are equivalent to the numerators of Wright's correlation coefficients [1, 68]. The majority of genetic connectedness literature is based on the pedigree relationship matrix, i.e., average relationships assuming conceptually, an infinite number of loci. On the other hand, the genomic relationship matrix, $\mathbf{K} = \mathbf{G}$, captures genomic similarity among individuals. The matrix $\mathbf{G}$ is a function of the matrix of allelic counts ($w_{i,j} \in 0, 1, 2$), where $i = 1, \cdots, n$ and $j = 1, \cdots, m$ denote the indices of individuals and of markers, respectively. Each element of the allele content matrix $\mathbf{W}$ is the number of copies of the reference allele. Under Hardy-Weinberg equilibrium, $E(w_{\cdot j}) = 2p_j$ and $Var(w_{\cdot j}) = 2p_j(1-p_j)$, so that $\mathbf{W}_{\cdot j} = \frac{w_{\cdot j} - 2p_j}{\sqrt{2p_j(1-p_j)}}$ is a standardized incidence matrix of allelic counts, where p_j is the allele frequency at the jth marker. The $\mathbf{G}$ matrix is constructed from a crossproduct of scaled marker genotype matrix $\mathbf{W}$ divided by some constant, i.e., the number of markers under assumption of unity marker variance

$$\mathbf{G} = \frac{\mathbf{W}\mathbf{W}'}{m}.$$

The standardization of $\mathbf{W}$ and the constant in the denominator make the $\mathbf{G}$ matrix analogous to the $\mathbf{A}$ matrix [69]. This genomic relationship matrix estimates the proportion of the genomes of two individuals that is identical by state (IBS).

One concern that arises when comparing the $\mathbf{A}$ and $\mathbf{G}$ matrices is that these two matrices are not on the same scale. The $\mathbf{G}$ matrix represents the estimate of a covariance (correlation) structure among individuals marked by SNP with the potential having some negative off-diagonal entries. Such negative values indicate that some individuals are molecularly less related than average pairs of individuals in the sense of IBS if the population were in Hardy-Weinberg equilibrium [e.g., 70]. This mostly happens when the current population is defined as a base population, namely, computing the $\mathbf{G}$ matrix by using the estimates of observed allele frequencies from the current population [71]. While the negative coefficients arising from IBS can be interpreted as negative correlations of alleles [70], this is contrast to the $\mathbf{A}$

matrix which is defined as an IBD. In the $\mathbf{A}$ matrix, a founder population is assumed to be the unselected base population. This may impact some of the connectedness statistics used in this study. For this reason, we also considered two other genomic relationship matrices: a $\mathbf{G_{0.5}}$ matrix and a scaled $\mathbf{G}$ matrix, $\mathbf{G_s}$, so that the genomic relationship matrix is on nearly the same scale as the $\mathbf{A}$ matrix. The $\mathbf{G_{0.5}}$ matrix was created by scaling the $\mathbf{W}$ by p_j^*, instead of p_j, where p_j^* is the estimate of allele frequency in the base population. Because allele frequencies in the base population are unknown, we set all p_j^* equal to 0.5 under the assumption of no selection [72, 73, 74]. The $\mathbf{G_{0.5}}$ matrix constructed in this way does not create any negative coefficients for the both mice and cattle datasets. The correlations between $\mathbf{G}$ and $\mathbf{G_{0.5}}$ (defined as correlation between elements of upper triangular matrix including diagonals) were 0.81 and 0.98 for mice and cattle, respectively.

Alternatively, a min-max scaler, one of the common scaling methods, was employed to scale the $\mathbf{G}$ matrix. The min-max scaler transforms inputs into the given range of minimum and maximum values. The scaled genomic relationship between ith and jth individual was given by

$$\mathbf{G_{s_{ij}}} = \frac{(\mathbf{G_{s}}_{max} - \mathbf{G_{s}}_{min})(\mathbf{G}_{ij} - \mathbf{G}_{min})}{\mathbf{G}_{max} - \mathbf{G}_{min}},$$

where $\mathbf{G}_{min}$ and $\mathbf{G}_{max}$ are the minimum and maximum elements of $\mathbf{G}$, and $\mathbf{G}_{ij}$ is the ith, jth element of $\mathbf{G}$. The $\mathbf{G_{s}}_{min}$ and $\mathbf{G_{s}}_{max}$ define the range of minimum and maximum values of elements of $\mathbf{G_s}$. These values were set to 0 and 2, respectively, according to the minimum and maximum values of numerators of Wright's correlation coefficients. This scaling sets negative off-diagonal entries in the $\mathbf{G}$ matrix to 0 [75]. Note that the correlation between $\mathbf{G}$ and $\mathbf{G_s}$ is equal to one because a correlation is invariant to changes in scale.

Lastly, the covariance between ungenotyped and genotyped individuals was jointly modeled

through a hyrbrid matrix where $\mathbf{K} = \mathbf{H}$. The $\mathbf{H}$ matrix can be viewed as a matrix that combines pedigree and genomic relationships. By considering the distribution of genetic values of ungenotyped individuals conditioned on genetic values of genotyped individuals, it can be shown [76, 77] that

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G}_{22} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}_{22} \\ \mathbf{G}_{22}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}_{22} \end{bmatrix}$$

where $\mathbf{A}_{11}$, $\mathbf{A}_{12}(\mathbf{A}_{21})$, and $\mathbf{A}_{22}$ are numerator relationship matrices among ungenotyped, ungenotyped and genotyped, and genotyped individuals, respectively. $\mathbf{G}_{22} = \mathbf{G}, \mathbf{G_{0.5}}$, or $\mathbf{G_s}$ is the genomic relationship matrix for genotyped individuals. In addition to $\mathbf{A}$, $\mathbf{G}$, $\mathbf{G_{0.5}}$, and $\mathbf{G_s}$, the $\mathbf{H}$ matrix was used for the cattle dataset that spans several generations. We treated individuals at generations three, four, and five as genotyped individuals and earlier generations as ungenotyped individuals. This reflects a practical situation in typical breeding programs, where the majority of genotyped individuals are concentrated in more recent generations. This partitioning resulted in 65% ungenotyped and 35% genotyped individuals, simulating a realistic scenario where there are more ungenotyped than genotyped individuals [e.g., 76].

3.3.7 Principal component analysis of measures of connectedness

Principal component analysis (PCA) of PEVD, CD, and r pairwise individual-based matrices computed under the four different simulated scenarios in the cattle dataset was used to cluster individuals. The prcomp function in R was used to produce principal component scores and the principal component (PC) plots were generated with the ggbiplot package based on the first two PC.

3.3.8 Heritability

For simulation, we used two heritability values ($h^2 = 0.8$ and $h^2 = 0.2$) by varying the ratio of variance components $\lambda = \sigma^2_\epsilon/\sigma^2_u = (1 - h^2)/h^2$ assuming an animal model, where σ^2_ϵ and σ^2_u are residual and genetic variances, respectively.

3.3.9 Data availability

The mouse dataset is available at http://wp.cs.ucl.ac.uk/outbredmice/heterogeneous-stock-mice/ and the cattle dataset is downloadable from the synbreedData R package at https://cran.r-project.org/web/packages/synbreedData/index.html.

3.4 Results

3.4.1 Mice data

Absence of full-sibs

The average (standard deviation) of pedigree relationships among individuals in the same management units was 0.491 (0.058) because of the aforementioned full-sib family assignments. The genomic counterpart **G**) gave a similar estimate of 0.494 with a slightly increased standard deviation of 0.087 due to Mendelian sampling variation [59]. The average across management unit pedigree-based genetic connectedness was 1.299 when measured by PEVD and $h^2 = 0.8$ (Table 3.1). Measures of connectedness increased using genomic data (**G**) by reducing PEVD to 0.456. With $h^2 = 0.2$, while the overall genetic connectedness decreased, genomic information (**G**) lowered PEVD compared to that of pedigree. Use of the $\mathbf{G_{0.5}}$

reduced PEVD more than that of the **G**, hence increased the measures of connectedness. Using the scaled genomic relationship matrix increased connectedness statistics compared to those of the pedigree-based, but they were less than those with **G**. Similarly, use of **G** matrix compared to the **A** matrix strengthened measured connectedness in CD for both $h^2 = 0.8$ and $h^2 = 0.2$. The $\mathbf{G_{0.5}}$ matrix also increased measures of connectedness compared to those of the **A** and the $\mathbf{G_s}$ matrix resulted in the greatest measures of connectedness among the four relatedness matrices. Both PEVD and CD statistics confirmed that genome-wide markers increased the degree of connectedness estimated between individuals across management units. However, the connectedness measures assessed by r were less when the **G** was compared with the **A**. On the other hand, the $\mathbf{G_{0.5}}$ and the scaled genomic relationship matrix $\mathbf{G_s}$ estimated greater connectedness measures than those of the **A**.

Presence of full-sibs

The increased estimates of disconnectedness were less when at least one full-sib was present in different management units for PEVD. For instance, comparisons between absence or presence of full-sibs across management units were 1.299 vs. 0.354 and 0.456 vs. 0.127 for pedigree-based vs. genome-based (**G**) PEVD, respectively. The presence of full-sibs in different management units decreased PEVD. However, corresponding statistics for CD were lower with the existence of full-sibs. This is explained by the fact that CD penalizes the estimates of connectedness when genetic variability is small. The CD statistic attempts to decrease the average PEV of the contrast while maintaining the variability of relatedness. Laloë [21] stated that increased estimate of connectedness should not be achieved by simply using genetically similar individuals and CD is the most relevant connectedness statistic in terms of genetic progress of agricultural species. This was confirmed in the mice data illustrating that the presence of full-sibs decreased the estimates of CD. Regardless of absence

or presence full-sibs across-units, genomic information elucidated additional relationships, thus increasing connectedness estimates relative to pedigree. This trend was also true for the $\mathbf{G_{0.5}}$ and $\mathbf{G_s}$ matrices. With r, when transitioning from $\mathbf{A}$ to $\mathbf{G}$, the values of the statistic reduced; however, the $\mathbf{G_{0.5}}$ and $\mathbf{G_s}$ yielded greater values of connectedness than those of pedigree in the existence of full-sibs. In all cases, using one of the $\mathbf{G}$, $\mathbf{G_{0.5}}$, or $\mathbf{G_s}$ matrix increased the estimates of connectedness statistics as compared to using the $\mathbf{A}$. As shown in Table 3.1, we found a similar overall pattern when h^2 was set to 0.2, although connectedness remained less than the alternative higher heritability. Replacing pedigree with genome-wide markers increased the degree of connectedness captured among individuals in disconnected management units.

Illustrative examples

To illustrate how $\mathbf{G}$ matrix impacted our measures of connectedness, we chose five management units including full-sib and non-full-sib individuals. In this example, management units "19F", "29A", and "36F" share at least one pair of full-sib individuals, whereas management units "12A" and "13C" do not share any full-sib individuals across management units. Figure 3.1 shows PEVD-derived connectedness across management units when $h^2 = 0.8$. Comparison across management units with full-sibs in common had smaller PEVD hence greater connectedness. The molecular information captures more of the genetic connectedness relative to pedigree for across management units. We further investigated how the $\mathbf{G}$ or $\mathbf{G_s}$ increased connectedness measures across management units relative to the $\mathbf{A}$ using PEVD and r. To do so we examined the specific components in the PEV matrix derived from several management units including full-sib and non-full-sib individuals. As shown in detail in Appendix B, we found that the rates of PEV (diagonals) and PEC (off-diagonals) reductions from $\mathbf{A}$ to $\mathbf{G}$ or $\mathbf{G_s}$ explain the changes of connectedness measures.

3.4.2 Cattle

Clustering

The partitioning around medoids clustering method yielded eight clusters. Table 3.2 contains descriptive statistics for those clusters. The number of individuals per cluster varied from 36 to 127. The average of within cluster pedigree-based relationships was around 0.05 except for cluster 6, in which distant relatives were grouped together. Between clusters, all pedigree-based relationships were close to zero. Each cluster was assigned to management units in four simulated scenarios as summarized in Figure 3.2.

- Scenario 1: Each cluster was assigned to its own management unit
- Scenario 2: Clusters 1, 2, 3, 4, and 5 were assigned to management unit 1 and clusters 6, 7, and 8 were assigned to management unit 2
- Scenario 3: Clusters 1, 2, 3 were assigned to management unit 1, clusters 7 and 8 were assigned to management unit 2, and individuals in clusters 4, 5, and 6 were assigned to both management units 1 and 2 to act as link among clusters or individuals that partially connect the two management units
- Scenario 4: Individuals in clusters 1 to 8 were equally assigned to management units 1 and 2

The number of individuals in management units 1 and 2 were approximately equal in scenarios 2, 3, and 4. We computed PEVD, CD, and r for each of the four scenarios and compared genetic connectedness when using the $\mathbf{A}$, $\mathbf{G}$, $\mathbf{G_s}$, and $\mathbf{H}$ kernel matrices.

Prediction error variance of the difference

Across management unit PEVD for each of four scenarios are presented in Table 3.3. Connectedness estimates increased across management units when transitioning from scenario 1 to scenario 4 for both heritability levels. Figure 3.3 shows the relative increase of genetic connectedness as measured with PEVD, as a percentage, across management units in comparison to scenario 1. Genetic connectedness across management units in scenario 1 was compared to across management unit connectedness obtained from scenarios 2, 3, and 4. We observed increased genetic connectedness as more individuals from the same clusters were shared between management units, resulting in the highest connectedness estimates in scenario 4. Transitioning from scenario 1 to scenario 4 increased connectedness for $\mathbf{A}$ and $\mathbf{G}$ for both heritability levels. The proportional increase in genetic connectedness in pedigree-based relationships were larger than those of genomic-based relationships because $\mathbf{G}$ matrix substantially increased measured connectedness between disconnected management units in scenario 1, reducing the gains in the following scenarios 2, 3 and 4. Also, as heritability increased, larger values of connectedness were observed. In general, with $\mathbf{G}$ and $\mathbf{G_{0.5}}$ increased the estimates of connectedness compared to those of the $\mathbf{A}$ regardless of heritability levels. This is in agreement with the mice dataset. However, with $\mathbf{G_s}$, values of PEVD were unexpected: although with scaled $\mathbf{G_s}$ produced estimates of connectedness that are higher than those with $\mathbf{A}$ when h^2 was set to 0.8, the same pattern was not observed for $h^2 = 0.2$.

Coefficient of determination

Across CD for each of four scenarios are presented in Table 3.3. Similar to PEVD the extent of connectedness across management units increased when moving from scenario 1 to scenario 2 and 3 regardless of the heritability levels. Figure 3.4 shows the percentage

increase in CD across management units when scenario 1 was treated as a base comparison. As with PEVD, CD statistics revealed an increase in the degree of connectedness as more individuals from the same clusters were assigned to different management units. However, the increase of CD was not observed when transitioning from scenario 1 to scenario 4. Again, this is because CD accounts for the reduction of connectedness due to reduced relatedness variability between individuals under comparison in scenario 4. This pattern was observed for both pedigree and genomic-based connectedness. Overall, $\mathbf{G}$, $\mathbf{G_{0.5}}$, and $\mathbf{G_s}$ all produced CD greater than those with $\mathbf{A}$ regardless of heritability level, yielding consistent measures of connectedness.

Prediction error correlation

Prediction error correlations across management units for each of four scenarios are presented in Tables 3.3. The results align with those of the mice dataset in that $\mathbf{G}$-based r statistics behave erratically in all scenarios making them difficult to interpret. However, the anticipated increases in r were observed with the transition from the $\mathbf{A}$ to the $\mathbf{G_{0.5}}$ or the scaled $\mathbf{G_s}$ matrix. Here $\mathbf{G_{0.5}}$ and $\mathbf{G_s}$-based measures consistently yielded greater connectedness values than those of pedigree counterparts. Figure 3.5 shows the percentage increases in r across management units when scenario 1 was treated as a base comparison. Here the $\mathbf{G_s}$ instead of the $\mathbf{G}$ matrix was used. The results align with those of PEVD and CD, where the extent of pedigree-based and genomic-based r statistics increase the most when more individuals from the same clusters were assigned to different management units. The magnitude of the increase was larger when heritability was greater. However, the increase of connectedness moving from Scenario 1 to 2 was not observed in pedigree-based measures. While both scenarios are not connected designs because pedigree-based relationships across the eight clusters were close to zero, it is interesting to note that with pedigree-based r scenario 2 was

more disconnected than scenario 1. From Kennedy and Trus [20], this is because stronger within unit connectedness can reduce between unit connectedness.

Ungenotyped and genotyped individuals

We considered a scenario where only individuals in younger generations were genotyped in the cattle dataset. For this purpose, we used the **H** matrix that blends ungenotyped and genotyped individuals. As shown in Table 3.4, results using the H matrix lie somewhere between the results obtained when using the **A**, **G**, and $\mathbf{G_s}$ matrices. This is expected because the **H** matrix was created from a combination of **A** and **G** or $\mathbf{G_s}$. Although an increase in measures of connectedness was observed compared to using the pedigree alone, this increase was smaller than when all individuals were genotyped. This finding suggests the possibility of strengthening the degree of connectedness even when only a subset of individuals was genotyped. An exception was observed when **H** constructed from $\mathbf{G_s}$ for PEVD: in this case the measures of connectedness were less than that from **A**.

Principal component analysis of connectedness

Principal component plots for CD derived from **A** and **G** matrices for scenarios 1 and 4 are presented in Figures 3.6 and 3.7, respectively. These correspond to the two extreme scenarios considered in the cattle dataset. In scenario 1 with $h^2 = 0.8$, eight clusters assigned to distinctive management units were separated from each other as expected using pedigree-based relationships (Figure 3.6). Genomic information brought these eight clusters closer to each other, thus shortening the distance between individuals from different management units. While eight clusters were less distinguishable from one another due to lower heritability, the same pattern was observed when h^2 was 0.2. These findings align with the fact that use

of genomic information increases measures of connectedness compared to pedigree. In both cases, cluster 6, which consisted of unrelated individuals, was clustered far away from the other clusters in the pedigree-based analysis. PCA yielded two clear clusters in scenario 4 when $h^2 = 0.8$, which correspond to the two management units considered (Figure 3.7). Replacing $\mathbf{A}$ with $\mathbf{G}$ resulted in a tighter concentration of a single cluster. A similar tendency was observed when $h^2 = 0.2$ supporting the findings that the extent of measures of connectedness between individuals from different management units is enhanced with genomic information. The remaining PC plots for PEVD ($\mathbf{A}$ and $\mathbf{G}$) and r ($\mathbf{A}$ and $\mathbf{G_s}$) are in Figures 3.8, 3.9, 3.10, and 3.11, which presented a pattern similar to Figures 3.6 and 3.7.

3.5 Discussion

With sufficient connectedness across management units, BLUP of genetic values can be fairly compared. Without such connectedness, making selection decisions based on breeding values of individuals from different management units might be associated with an increased risk of uncertainty in genetic evaluation due to imperfect separation of the genetic signal from noise. In addition to PEVD, CD, and r, other connectedness measures have been applied to pedigree data [e.g., 78, 79], which have their own characteristics. Advancement of molecular biotechnology now enables us to assess connectedness at the genomic level. Although genomic data are clearly important in genetic evaluations due to increased accuracy of estimates of genetic merit for non-parent individuals, little consideration has been given to the effect of genomic information on connectedness measures. In this study, we employed three measures of connectedness to examine the extent to which genomic information increases the estimates of connectedness.

Relatedness in quantitative genetics

The majority of connectedness among management units was driven by the degree of genetic links or relatedness. The theory behind relatedness is largely entrenched in quantitative genetics dating back to work of Fisher [80] and Wright [1]. Quantitative genetics offers a useful framework to study traits and diseases that are controlled by a considerable number of small effect genes. For traits with polygenic genetic architectures, inheritance does not exhibit a clear genotype-phenotype pattern. However, genetic resemblance between relatives (e.g., the genetic correlations between parent and offspring or between pairs of different types of siblings) can be exploited to estimate quantitative genetic parameters. For this reason, genetic resemblance between relatives has been at the heart of quantitative genetics. Consequently, the vast majority of the theoretical developments and applications of the last century were built around family data. The availability of dense panels of common SNP has made it possible to trace Mendelian sampling and hence capture more detailed relatedness compared to pedigree information. It enables quantifying genomic kinships among related individuals that are not otherwise apparent because of incomplete pedigrees or the general assumption that animals in a baseline or founder population are unrelated. Thus, it has opened new opportunities for quantitative genetic analysis using data from distant relatives. The rationale is that individuals are genomically related to some extent and molecular similarity introduces covariance even if individuals are not related in the sense of known pedigree. These factors possibly contribute to the reduction of PEV or increase of PEC and hence lead to increased capturing of genetic connectedness in PEVD, CD, and r such that genetic merit estimates can be better compared across management units.

The impact of genomic information on connectedness

We found from the mice data that genomic information increased favorable changes in measures of connectedness among individuals from different management units and reduced the risk of potential uncertainty in EBV-based comparisons when selecting individuals across management units. In addition, the rate of improvement in measures of connectedness in PEVD and r was greater when there was at least one full-sib in different management units. This is in concordance with Legarra et al. [62] who used the same dataset and reported that the use of genome-wide selection increased predictive performance up to 0.22 across families and up to 0.03 within families compared to pedigree-based regression counterparts. On the other hand, CD accounted for the reduction of variability of relatedness between individuals under comparison resulting in decreased estimates of connectedness. Analysis of cattle data supported the results from mice and revealed that the benefit of using genomic information is greater for a disconnected design rather than a connected design. PCA was performed to visualize improvement in connectedness when moving from pedigree to genomic-based relationships. The PC plots supported the evidence that genomic information can improve detection of connectedness between individuals from different management units. This is particularly so when more individuals from the same clusters are assigned to different management units.

Choice of kernel matrices

Unlike PEVD and CD, comparisons between the **A** and **G** kernel matrices evaluated by the r statistic behaved irregularly. By examining the specific components of the PEV matrix for **G** and **A** in the mice dataset, we found genomic information reduces off-diagonal elements more than diagonals. This illustrates a fundamental difference between r and either PEVD or CD

because this statistic is based on the ratio, rather than the magnitude of individual elements. It may be argued that the inconsistent connectedness results from r occur because the $\mathbf{G}$ matrix is not on the same scale as the $\mathbf{A}$ matrix, suggesting that r statistics are not invariant regarding how genomic relatedness is defined. Given pedigree information, the numerator relationship matrix is defined as IBD. On the other hand, given a marker matrix, there are a number of ways to construct a genomic relationship matrix as discussed by Toro et al. [73]. The $\mathbf{G}$ matrix we used captures the proportion of the genome that is IBS by accounting for the covariance structure among individuals by molecular markers [70]. This kernel matrix is an estimator of IBD relationships [71]. Caution should be exercised when interpreting connectedness measures derived using genomic data as the underlying assumption is that relationships are built based on alleles being IBS and not necessarily being IBD. Therefore, we attempted to make $\mathbf{G}$ more compatible to $\mathbf{A}$ by using $\mathbf{G_{0.5}}$ derived from allele frequencies equal to 0.5 and by using the min-max scaler transformation to produce the scaled genomic relationship matrix $\mathbf{G_s}$. For instance, compared to using $\mathbf{G}$, entries of PEV matrix from using $\mathbf{G_s}$ were more similar to those $\mathbf{A}$, especially when there was connectedness, and in turn r statistics yielded greater connectedness values. Although connecting marker-based genomic relatedness to classical theory is still an open question in quantitative genetics, care needs to be taken when comparing genetic connectedness with genomic connectedness especially when the ratio-based statistic is used. Moreover, many additional factors may influence the elements of IBS matrix such as the choice of MAF, the density of SNP, imperfect linkage disequilibrium (LD) between markers and quantitative trait loci, and error associated with estimating genomic relationships from a finite set of markers [e.g., 81].

Choice of connectedness statistics

There was an issue with PEVD coupled with the $\mathbf{G_s}$ matrix in the cattle dataset when h^2 was 0.2 as the estimates of connectedness were less than those using $\mathbf{A}$ (Table 3.3). Note that this was not the case when h^2 was equal to 0.8. The $\mathbf{H}$ matrix blended from the $\mathbf{A}$ and $\mathbf{G}$ kernel matrices yielded the estimates of connectedness that lie somewhere between the results obtained when using $\mathbf{A}$ and $\mathbf{G}$ alone. However, this pattern was not observed when $\mathbf{G_s}$ was used in conjunction with $\mathbf{A}$ to compute PEVD (Table 3.4). Apparently, scaling has a negative influence on blending for PEVD, which warrants further research. One potential reason with $\mathbf{G_s}$ for the discrepancy is the proportional increase of PEC relative to PEV is larger when transitioning from the $\mathbf{A}$ to $\mathbf{G_s}$. This issue of proportional change is similar to that observed earlier with the r statistic coupled with $\mathbf{G}$. These results illustrate that connectedness statistics are not invariant with respect to how the genomic relationship matrix is created and each of them captures different aspects of genomic connectedness. The CD was the only one statistic that yielded consistent estimates of increased connectedness throughout this study. Its consistency was observed regardless of choice of kernel matrices, heritability levels, datasets used, and simulated scenarios for management units.

Inferring variance components from data

One concern with the current study is fixing heritability levels for all scenarios based on the assumption that both pedigree and genomic relationship matrices explain the equal amount of heritability. In practice, this assumption might not hold true when SNP do not capture the entire QTL signals. To address this concern, additional analysis was carried out such that variance components were estimated from the data rather than assuming these were known. We analyzed two publicly available phenotypes in the synbreedData R package

(Wimmer et al., 2015) for the cattle data used in this study. The heritabilities of these traits are 0.66 and 0.41. Connectedness analysis in Table 3.3 was repeated based on variance components estimated by a restricted maximum likelihood. The measures of CD derived from the A and G matrices are shown in Table 3.5. We found that genomic relatedness also increased connectedness measures more so than those of pedigree when variance components were directly estimated from the data. This result was consistent with what we found in the cattle data analysis reported in Table 3.3.

Future direction

One important direction for future study is to investigate whether increased connectedness observed by genomic relatedness also leads to increased predictive accuracy of genetic values across management units assessed by cross-validation. In this case, across management units can be considered as training and testing sets. In addition, while the current norm of genomic prediction is to use an IBS relationship matrix that aims to capture relationships at unknown QTLs through LD between markers and QTLs, we argue that improving the quality of breeding value comparisons and improving the accuracy of genomic prediction can be viewed as relevant but two different items. In this regard, a genome-wide IBD relationship matrix [e.g., 82], where marker inheritance is traced through a known pedigree, may be worthwhile to revisit for the purpose of ascertaining connectedness in a future study.

Also, for the r statistic, we summarized connectedness by averaging the r statistic of pairs of individuals across units rather than by averaging the relevant components of PEC and PEV followed by taking their ratio; our justification for that choice was provided in the Appendix A. When the latter summary statistic was used for r, the differences were negligible in the mice data and the pattern was the same for the scenario 1 of cattle data.

In conclusion, this study confirms that use of genomic relatedness improved genetic connectedness across management units compared to use of pedigree relationships. To our knowledge, this marks the first thorough investigation of genomic connectedness. We contend that our work is a critical first step toward better understanding genetic connectedness that may have a positive impact on genomic evaluation of agricultural species.

3.6 Figures

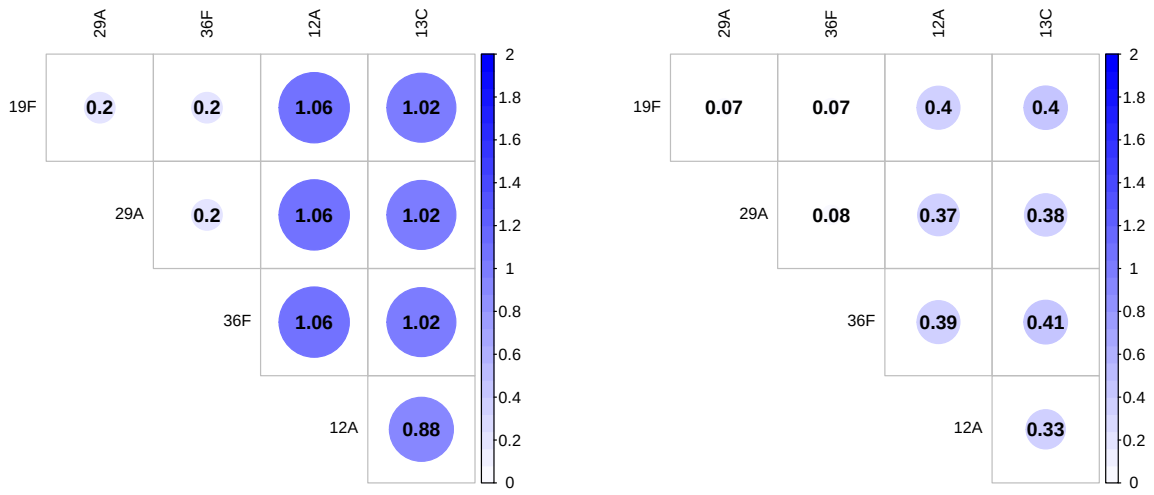

Figure 3.1: Prediction error variance of the difference (PEVD) for across five management units in the mice dataset. Management units "19F", "29A", and "36F" share at least one pair of full-sibs individudals with each other, whereas "12A" and "13C" do not share any individuals across management units. The left and right are pedigree-based (**A**) and genomic-based (**G**) connectedness, respectively. Darker color represents less genetic connectedness.

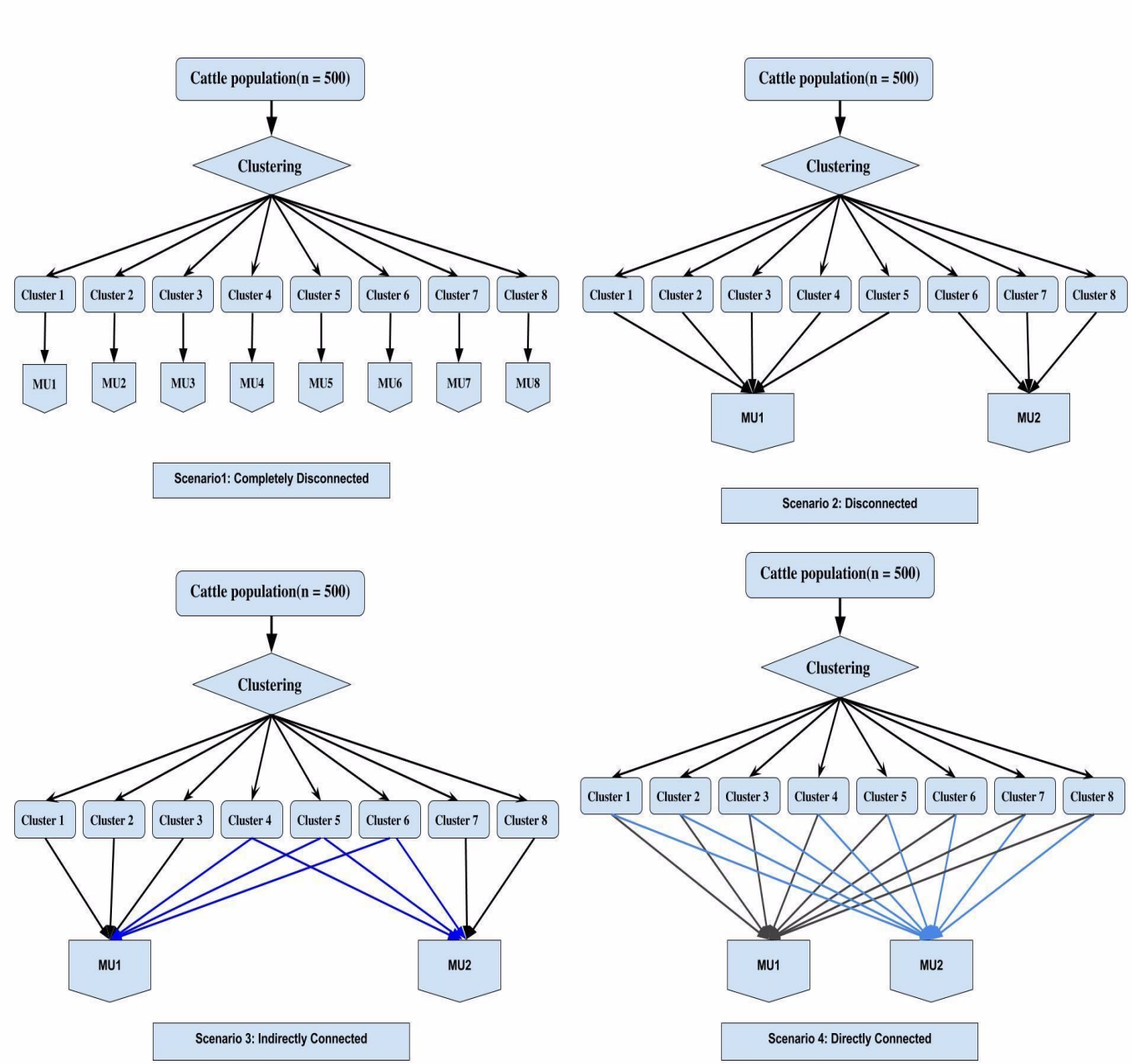

Figure 3.2: Four simulation scenarios considered in the cattle dataset. MU stands for management unit. Scenario 1: Completely disconnected - 8 clusters assigned to separate management unit. Scenario 2: Disconnected - clusters 1, 2, 3, 4, and 5 assigned to management unit 1 and clusters 6, 7, and 8 assigned to management unit 2. Scenario 3: Partially connected - clusters 1, 2, 3 assigned to management unit 1, clusters 7 and 8 assigned to management unit 2, and the remaining clusters 4, 5, and 6 assigned to both management units 1 and 2 that act as link among clusters or individuals that partially connect the two management units. Scenario 4: Connected - all clusters 1 to 8 were equally assigned to management units 1 and 2.

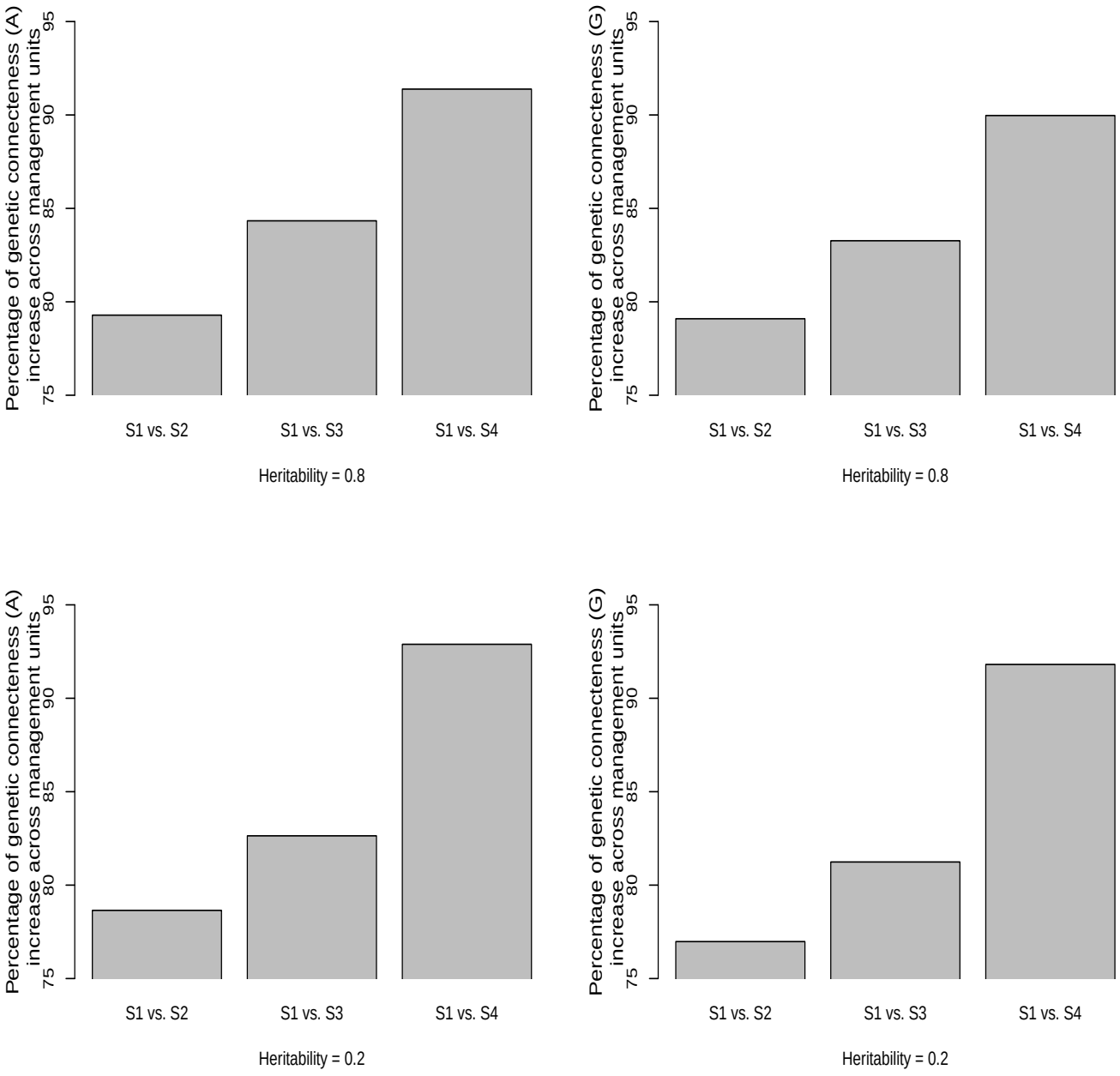

Figure 3.3: Percentage of relative increase in prediction error variance of the difference (PEVD) across management units in comparison to base Scenario 1. Two heritability values 0.8 and 0.2 were simulated. S1 (completely disconnected), S2 (disconnected), S3 (partially connected), and S4 (connected) represent four management unit scenarios. Left: **A** matrix. Right: **G** matrix.

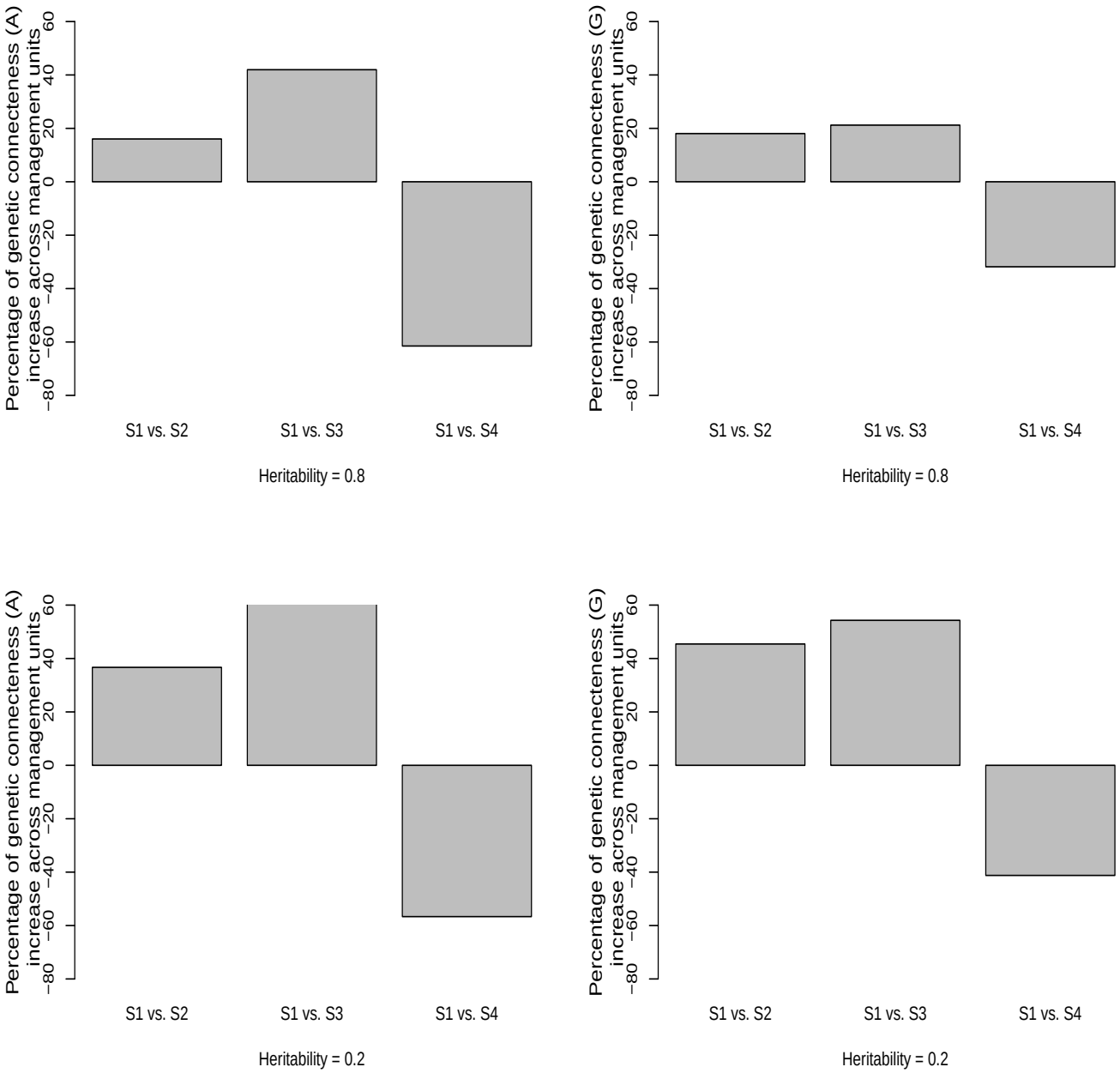

Figure 3.4: Percentages of relative increase in coefficient of determination of the difference (CD) across management units in comparison to base scenario 1. Two heritability values 0.8 and 0.2 were simulated. S1 (completely disconnected), S2 (disconnected), S3 (partially connected), and S4 (connected) represent four management unit scenarios. Left: **A** matrix. Right: **G** matrix.

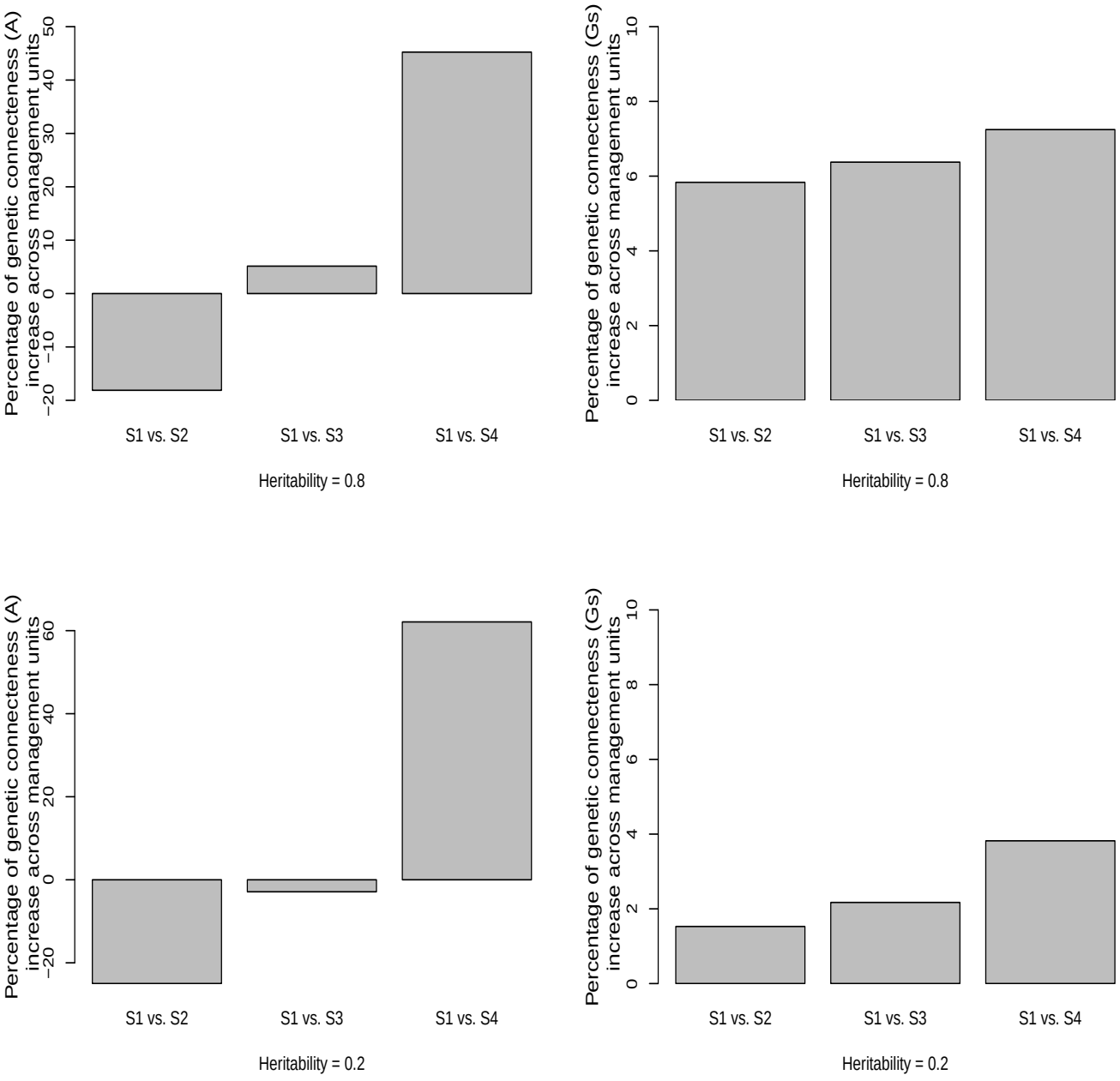

Figure 3.5: Percentages of relative increase in prediction error correlation (r) across management units in comparison to base scenario 1. Two heritability values 0.8 and 0.2 were simulated. S1 (completely disconnected), S2 (disconnected), S3 (partially connected), and S4 (connected) represent four management unit scenarios. Left: **A** matrix. Right: **Gs** matrix.

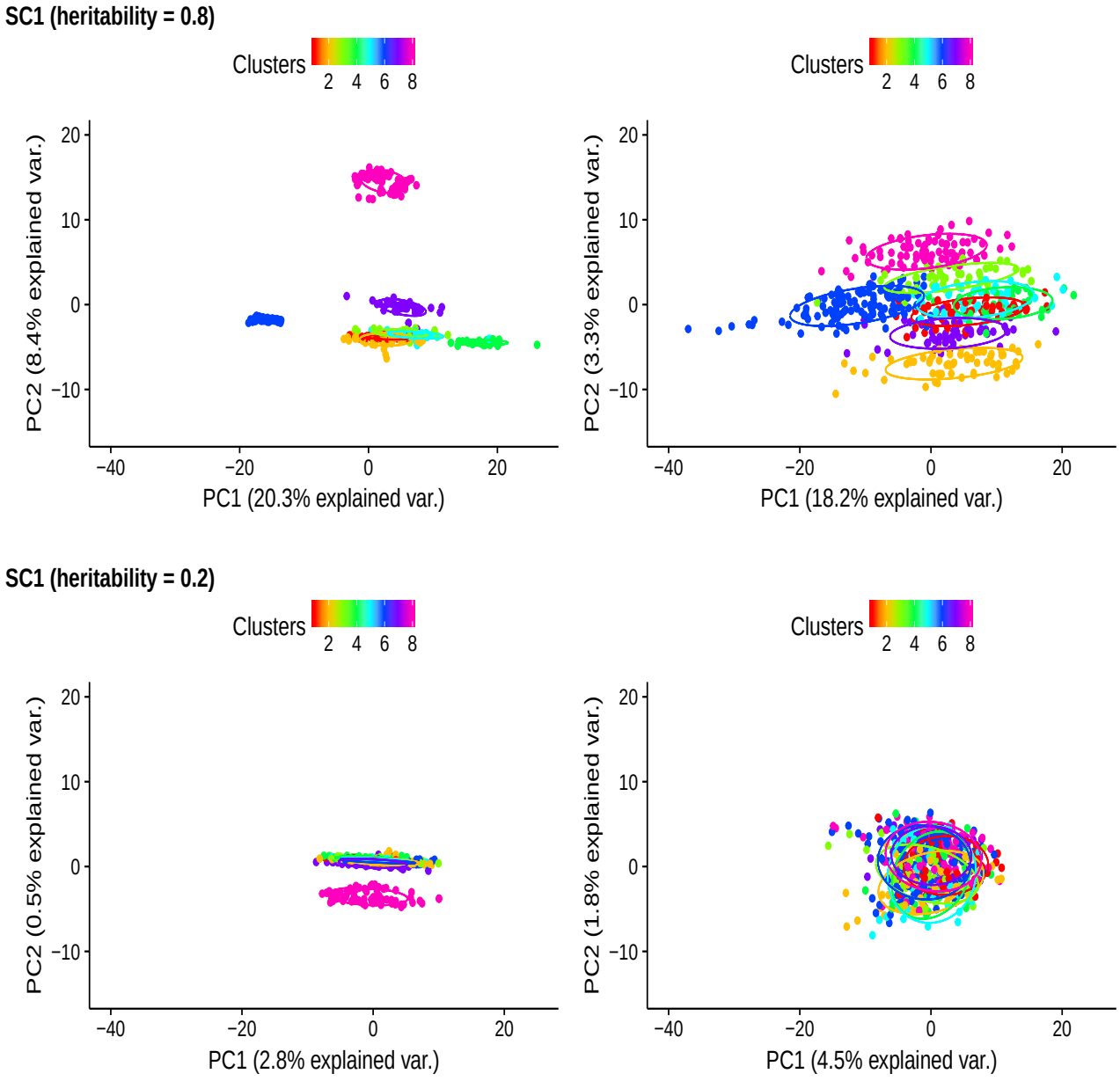

Figure 3.6: Principal component (PC) plots for Scenario 1 with coefficient of determination (CD) statistics. The first and second rows are according to heritability of 0.8 and of 0.2. The first and second columns are derived from pedigree-based (**A**) and genome-based (**G**) CD, respectively. The PC plots were grouped by clusters and colored in different colors. Individuals within the same cluster were grouped by the circles.

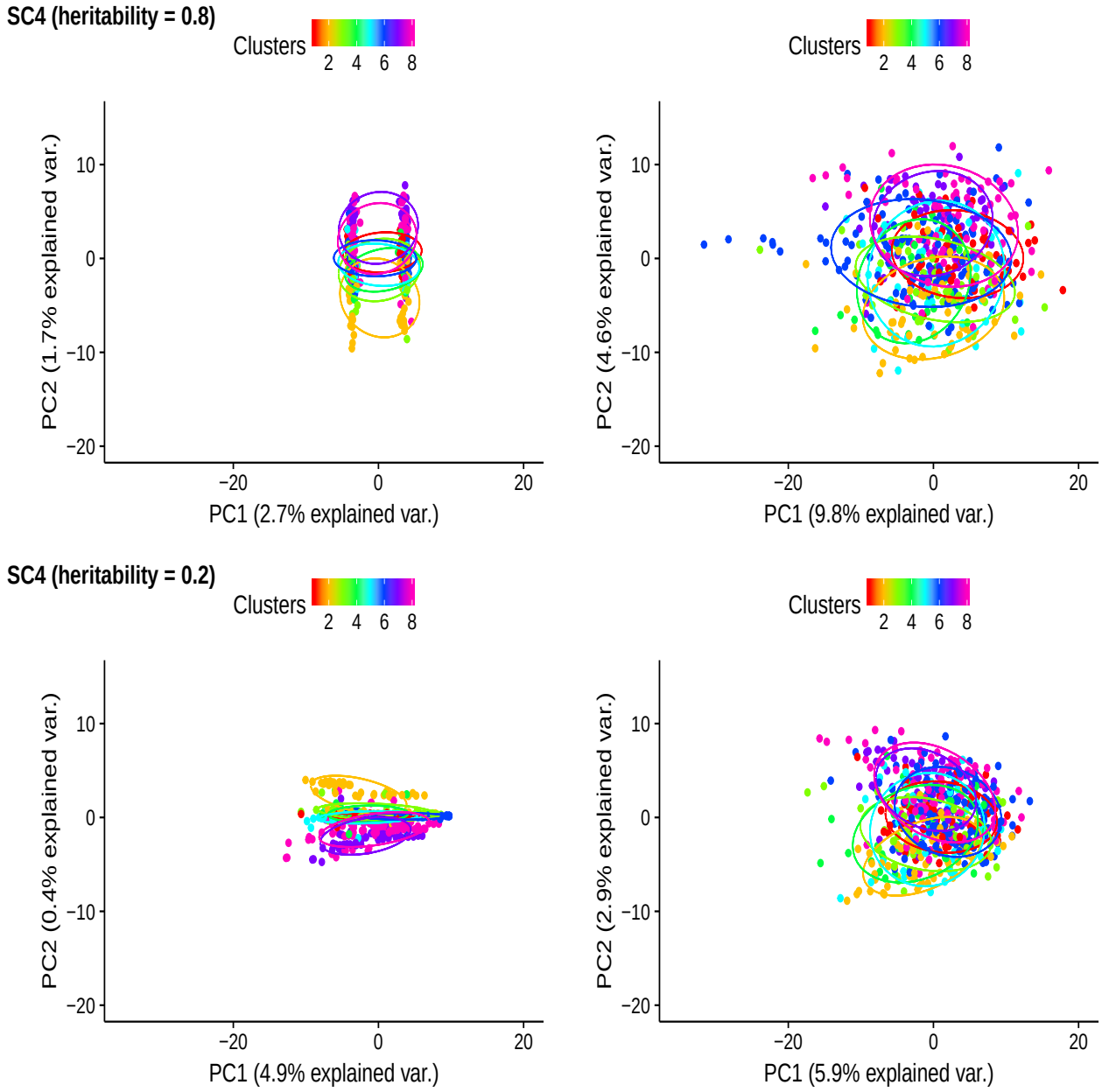

Figure 3.7: Principal component (PC) plots for Scenario 4 with coefficient of determination (CD) statistics. The first and second rows are according to heritability of 0.8 and of 0.2. The first and second columns are derived from pedigree-based (**A**) and genome-based (**G**) CD, respectively. The PC plots were grouped by clusters and colored in different colors. Individuals within the same cluster were grouped by the circles.

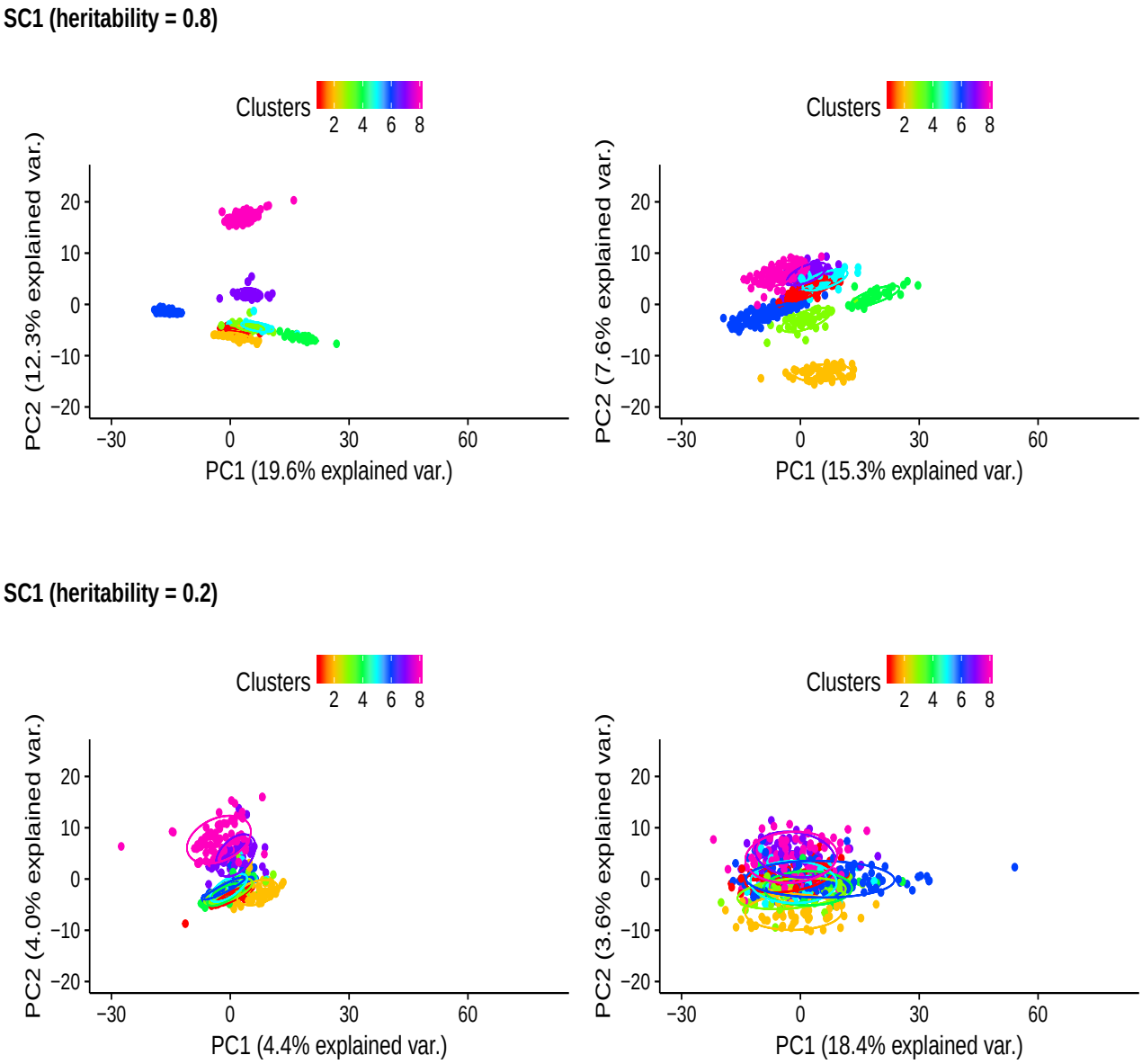

Figure 3.8: Principle component (PC) plots for Scenario 1 with prediction error variance of the difference (PEVD) statistics. The first and second rows are according to heritability of 0.8 and of 0.2. The first and second columns are derived from pedigree-based and genome-based PEVD, respectively. The PC plots were grouped by clusters and colored in different colors. Individuals within the same cluster were grouped by the circles.

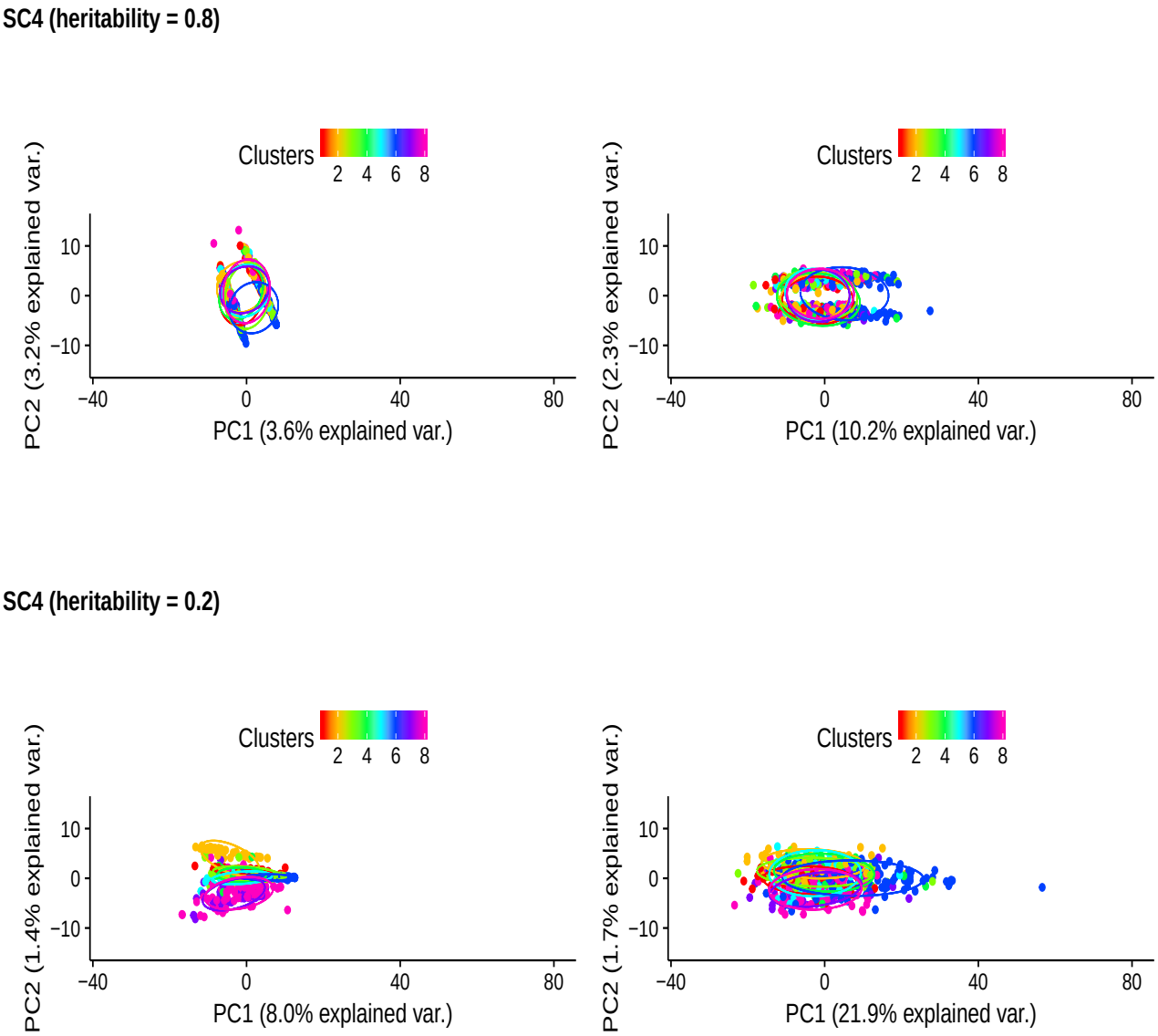

Figure 3.9: Principle component (PC) plots for Scenario 4 with prediction error variance of the difference (PEVD) statistics. The first and second rows are according to heritability of 0.8 and of 0.2. The first and second columns are derived from pedigree-based and genome-based PEVD, respectively. The PC plots were grouped by clusters and colored in different colors. Individuals within the same cluster were grouped by the circles.

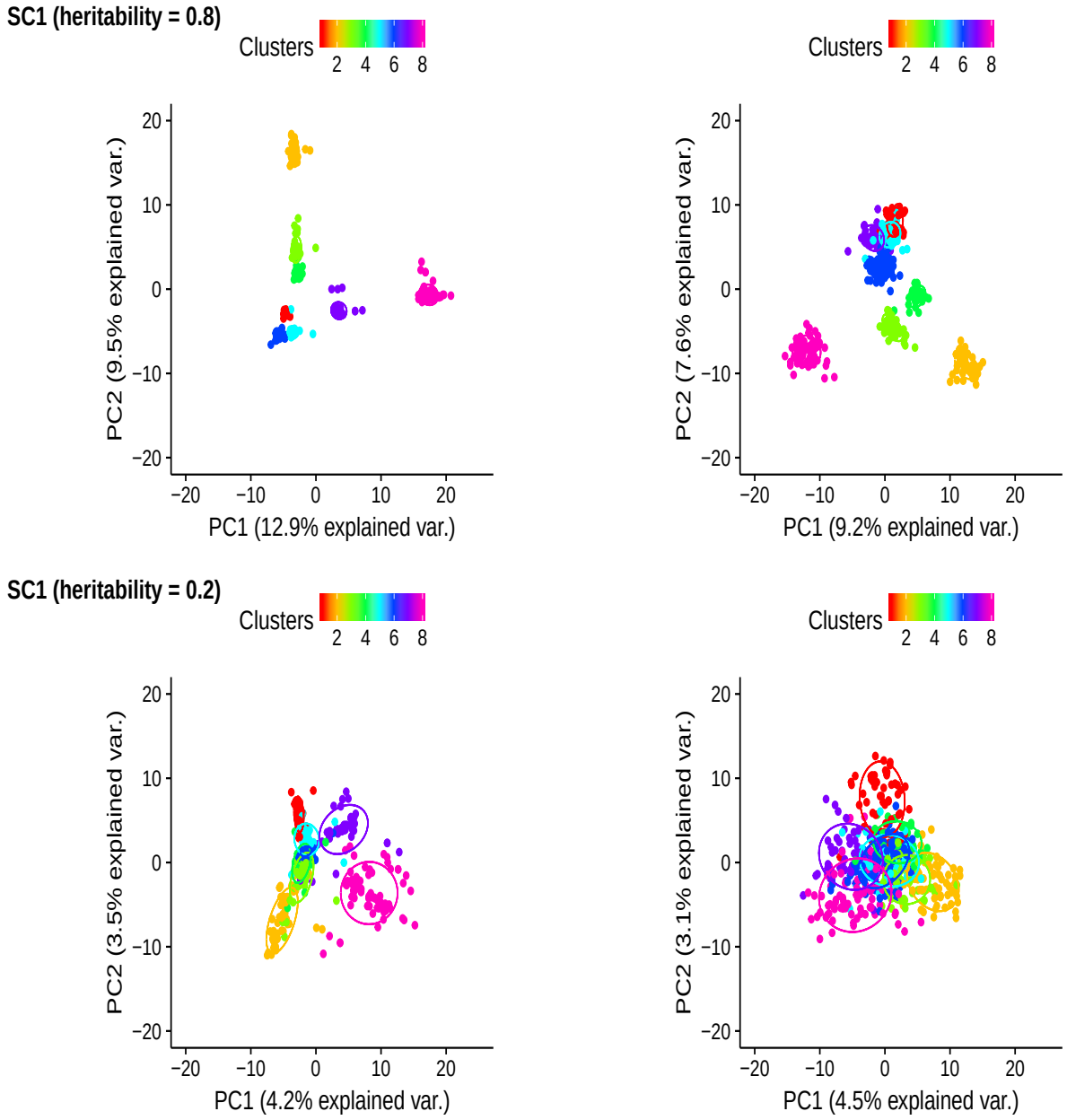

Figure 3.10: Principle component (PC) plots for Scenario 1 with prediction error correlation (r) statistics. The first and second rows are according to heritability of 0.8 and of 0.2. The first and second columns are derived from pedigree-based and genome-based r, respectively. The PC plots were grouped by clusters and colored in different colors. Individuals within the same cluster were grouped by the circles.

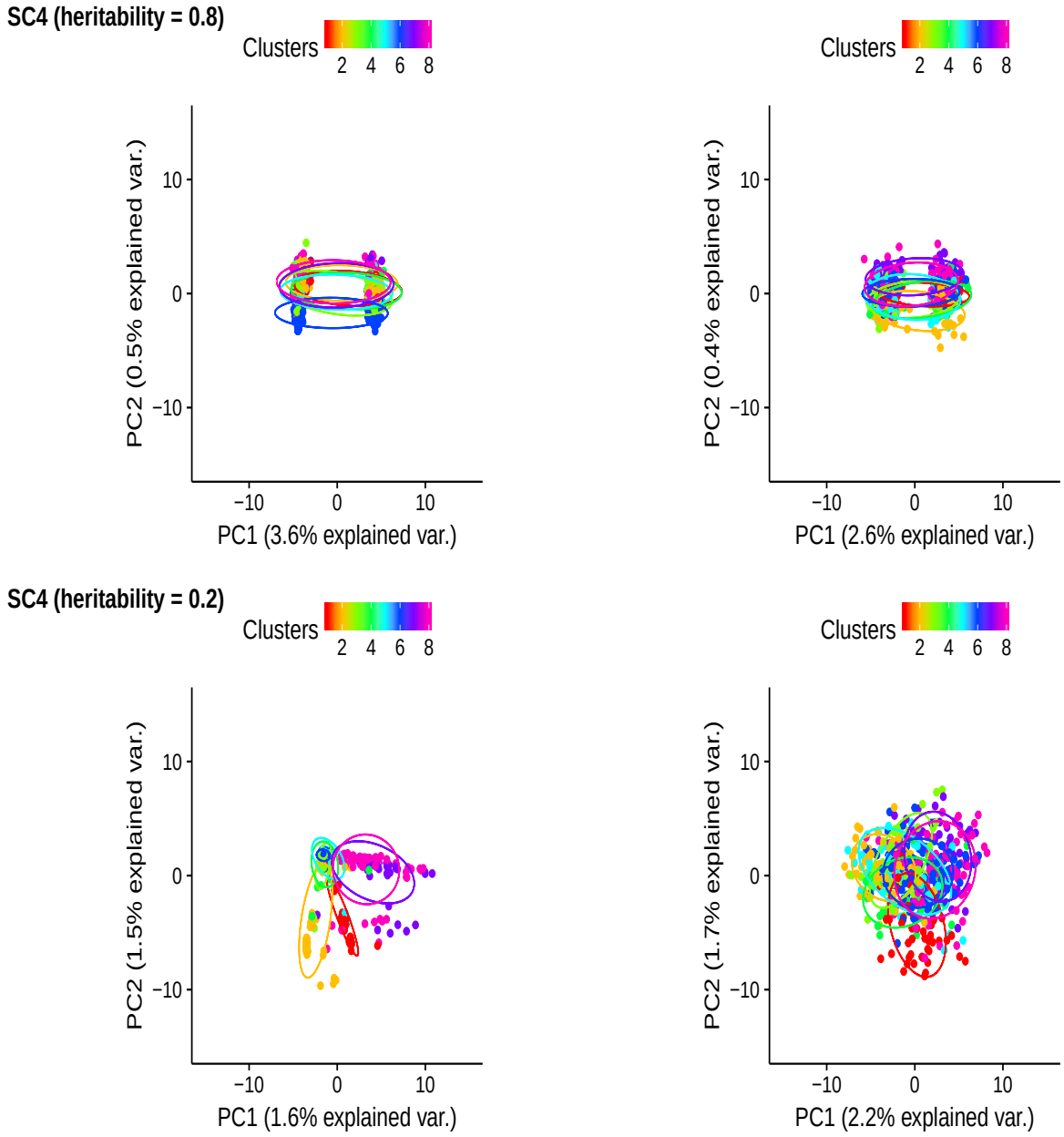

Figure 3.11: Principle component (PC) plots for Scenario 4 with prediction error correlation (r) statistics. The first and second rows are according to heritability of 0.8 and of 0.2. The first and second columns are derived from pedigree-based and genome-based r, respectively. The PC plots were grouped by clusters and colored in different colors. Individuals within the same cluster were grouped by the circles.

3.7 Tables

Table 3.1: Average genetic connectedness measures across management units in the mice data. PEVD, CD, and r denote prediction error variance of the difference, coefficient of determination, and prediction error correlation. We compared pedigree-based **A**, standard genome-based **G**, genome-based $\mathbf{G_{0.5}}$ assuming equal allele frequencies, and scaled genome-based $\mathbf{G_s}$ matrices to evaluate relationships among individuals. Two heritability values 0.8 and 0.2 were simulated. Values inside parentheses represent connectedness when at least one full-sib pair was present in different management units.

		Heritability (h^2)	
Methods	Kernels	0.8	0.2
PEVD	**A**	1.299 (0.354)	1.331 (0.366)
	G	0.456 (0.127)	1.037 (0.285)
	$\mathbf{G_{0.5}}$	0.374 (0.104)	0.824 (0.224)
	$\mathbf{G_s}$	0.532 (0.149)	1.254 (0.345)
CD	**A**	0.034 (0.024)	0.009 (0.007)
	G	0.662 (0.650)	0.234 (0.227)
	$\mathbf{G_{0.5}}$	0.640 (0.624)	0.207 (0.199)
	$\mathbf{G_s}$	0.690 (0.678)	0.270 (0.262)
r_{ij}	**A**	0.004 (0.600)	0.003 (0.486)
	G	-0.001 (0.525)	-0.001 (0.478)
	$\mathbf{G_{0.5}}$	0.559 (0.794)	0.433 (0.708)
	$\mathbf{G_s}$	0.496 (0.771)	0.270 (0.622)

Table 3.2: Descriptive statistics of the 8 clusters created by partitioning around medoids in the cattle data.

Cluster	Number of individuals	Average pedigree relationship
1	52	0.054
2	61	0.053
3	46	0.040
4	36	0.052
5	43	0.043
6	127	0.005
7	55	0.055
8	80	0.047

Table 3.3: Average genetic connectedness statistics across management units in the cattle data. S1 (completely disconnected), S2 (disconnected), S3 (partially connected), and S4 (connected) represent four management unit scenarios. PEVD, CD, and r denote prediction error variance of the difference, coefficient of determination, and prediction error correlation. We compared pedigree-based $\mathbf{A}$, standard genome-based $\mathbf{G}$, genome-based $\mathbf{G_{0.5}}$ assuming equal allele frequencies, and scaled genome-based $\mathbf{G_s}$ kernel matrices to evaluate relationships among individuals. Two heritability values 0.8 and 0.2 were simulated.

			Heritability (h^2)	
Scenarios	Methods	Kernels	0.8	0.2
S1	PEVD	$\mathbf{A}$	0.077	0.102
		$\mathbf{G}$	0.051	0.085
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.039 (0.066)	0.066 (0.110)
	CD	$\mathbf{A}$	0.324	0.112
		$\mathbf{G}$	0.539	0.224
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.528 (0.558)	0.195 (0.265)
	r_{ij}	$\mathbf{A}$	0.017	0.005
		$\mathbf{G}$	-0.014	-0.007
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.725 (0.468)	0.465 (0.174)
S2	PEVD	$\mathbf{A}$	0.016	0.022
		$\mathbf{G}$	0.011	0.020
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.008 (0.014)	0.015 (0.025)
	CD	$\mathbf{A}$	0.376	0.152
		$\mathbf{G}$	0.636	0.326
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.625 (0.652)	0.290 (0.373)
	r_{ij}	$\mathbf{A}$	0.014	0.004
		$\mathbf{G}$	-0.015	-0.007
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.738 (0.496)	0.468 (0.177)
S3	PEVD	$\mathbf{A}$	0.012	0.018
		$\mathbf{G}$	0.008	0.016
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.007 (0.011)	0.013 (0.020)
	CD	$\mathbf{A}$	0.460	0.211
		$\mathbf{G}$	0.653	0.346
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.649 (0.669)	0.312 (0.394)
	r_{ij}	$\mathbf{A}$	0.018	0.005
		$\mathbf{G}$	-0.012	-0.006
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.739 (0.498)	0.468 (0.178)
S4	PEVD	$\mathbf{A}$	0.007	0.007
		$\mathbf{G}$	0.005	0.007
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.004 (0.007)	0.005 (0.009)
	CD	$\mathbf{A}$	0.125	0.048
		$\mathbf{G}$	0.367	0.132
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.362 (0.384)	0.114 (0.158)
	r_{ij}	$\mathbf{A}$	0.024	0.008
		$\mathbf{G}$	-0.007	-0.002
		$\mathbf{G_{0.5}}$ ($\mathbf{G_s}$)	0.741 (0.502)	0.470 (0.181)

Table 3.4: Average genetic connectedness statistics across management units in the cattle data. S1 (completely disconnected), S2 (disconnected), S3 (partially connected), and S4 (connected) represent four management unit scenarios. PEVD, CD, and r denote prediction error variance of the difference, coefficient of determination, and prediction error correlation. We combined pedigree-based $\mathbf{A}$ with the standard genome-based $\mathbf{G}$, genome-based $\mathbf{G_{0.5}}$ assuming equal allele frequencies, and scaled genome-based $\mathbf{G_s}$ kernel matrices to evaluate relationships among individuals. Two heritability values 0.8 and 0.2 were simulated.

			Heritability (h^2)	
Scenarios	Methods	Kernels	0.8	0.2
S1	PEVD	$\mathbf{H}$ ($\mathbf{G}$)	0.069	0.095
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.089	0.130
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.062	0.090
	CD	$\mathbf{H}$ ($\mathbf{G}$)	0.379	0.140
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.473	0.232
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.504	0.283
	r_{ij}	$\mathbf{H}$ ($\mathbf{G}$)	-0.012	-0.004
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.054	0.036
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.066	0.060
S2	PEVD	$\mathbf{H}$ ($\mathbf{G}$)	0.014	0.021
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.019	0.028
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.014	0.020
	CD	$\mathbf{H}$ ($\mathbf{G}$)	0.464	0.216
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.530	0.287
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.531	0.306
	r_{ij}	$\mathbf{H}$ ($\mathbf{G}$)	-0.015	-0.004
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.040	0.025
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.050	0.041
S3	PEVD	$\mathbf{H}$ ($\mathbf{G}$)	0.011	0.017
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.013	0.021
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.010	0.015
	CD	$\mathbf{H}$ ($\mathbf{G}$)	0.510	0.251
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.589	0.344
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.521	0.265
	r_{ij}	$\mathbf{H}$ ($\mathbf{G}$)	-0.011	-0.003
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.042	0.026
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.051	0.041
S4	PEVD	$\mathbf{H}$ ($\mathbf{G}$)	0.007	0.007
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.008	0.009
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.006	0.007
	CD	$\mathbf{H}$ ($\mathbf{G}$)	0.168	0.060
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.255	0.126
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.256	0.145
	r_{ij}	$\mathbf{H}$ ($\mathbf{G}$)	-0.005	0.001
		$\mathbf{H}$ ($\mathbf{G_s}$)	0.049	0.029
		$\mathbf{H}$ ($\mathbf{G_{0.5}}$)	0.056	0.044

Table 3.5: Average genetic connectedness measured as coefficient of determination (CD) across management units in the cattle data. S1 (completely disconnected), S2 (disconnected), S3 (partially connected), and S4 (connected) represent four management unit scenarios. We compared pedigree-based **A** with the standard genome-based **G** kernel matrices to evaluate relationships among individuals. Two traits with heritability values of 0.66 (Trait 1) and 0.41 (Trait 2) were analyzed and variance components were estimated from the data rather than assumed known.

		Traits	
Scenarios	Kernels	Trait 1	Trait 2
S1	**A**	0.282	0.200
	G	0.345	0.266
S2	**A**	0.336	0.252
	G	0.457	0.374
S3	**A**	0.421	0.332
	G	0.480	0.395
S4	**A**	0.110	0.081
	G	0.213	0.159

Chapter 4

Do stronger measures of genomic connectedness enhance prediction accuracies across management units?

4.1 Abstract

Genetic connectedness assesses the extent to which estimated breeding values can be fairly compared across management units. Ranking of individuals across units based on best linear unbiased prediction (BLUP) is reliable when there is a sufficient level of connectedness due to a better disentangling of genetic signal from noise. Connectedness arises from genetic relationships among individuals. Although a recent study showed that genomic relatedness strengthens the estimates of connectedness across management units compared to that of pedigree, the relationship between connectedness measures and prediction accuracies only has been explored to a limited extent. In this study, we examined whether increased measures of connectedness led to higher prediction accuracies evaluated by a cross-validation based on computer simulations. We applied prediction error variance of the difference, coefficient of determination, and BLUP-type prediction models to data simulated under various scenarios. We found that a greater extent of connectedness enhanced accuracy of whole-genome prediction. The impact of genomics was more marked when large numbers of

markers were used to infer connectedness and evaluate prediction accuracy. Connectedness across units increased with the proportion of connecting individuals and this increase was associated with improved accuracy of prediction. The use of genomic information resulted in increased estimates of connectedness and improved prediction accuracies compared to those of pedigree-based models when there was enough markers to capture variation due to QTL signals.

4.2 Introduction

Genetic connectedness quantifies the extent of risk associated with the comparisons of estimated breeding values (EBV) across management units [9]. Best linear unbiased prediction (BLUP) of EBV can be fairly compared across units in the presence of a sufficient level of connectedness. On the other hand, an insufficient level of connectedness increases the risk of uncertainty in EBV comparisons when selecting individuals across units due to imperfect uncoupling of genetic signal from noise. A number of studies have shown that increasing pedigree-based connectedness through exchange of common reference sires can result in more accurate comparisons of genetic values of individuals from different management units [12, 13, 14]. The magnitude of estimates of connectedness is a function of genetic relatedness or relationships among individuals. Despite the critical importance of connectedness towards enabling genetic evaluations, the impact of genomic information on the degree of connectedness relative to pedigree only has been explored to a limited extent.

Use of genomics can affect genetic evaluations in two related but different contexts. One is related to determining if EBV can be safely compared across management units and the other is related to enhancing the reliability of EBV. In the former context, Yu et al. [83] employed three measures of connectedness to examine the extent to which genomic

information increases the estimates of connectedness. They found that the use of genomic relatedness improved genetic connectedness measures across management units compared to the use of pedigree relationships.

However, it remains an open question as to whether increased connectedness observed by genomic relatedness also leads to increased prediction accuracy of genetic values across management units. While improving the quality of breeding value comparisons and improving the accuracy of genomic prediction have been discussed in different contexts historically, it is worth investigating how these two items are related to each other. The objectives of this study were to examine how choice of relationship matrices and connectedness statistics impact the estimates of connectedness under various simulated scenarios and to assess the relationship between connectedness level and genome-enabled prediction accuracy. In addition, a guideline with respect to a sufficient level of connectedness is discussed.

4.3 Materials and Methods

4.3.1 Data simulation

Ten replicates of genotypes and phenotypes were simulated using the QMSim software [84] with details summarized in Figure 4.1. One single historical population with 1,100 generations was simulated with the forward-in-time approach to create the initial linkage disequilibrium (LD) and mutation-drift equilibrium. The mating system was based on the random union of gametes sampled from sires and dams and the only evolutionary forces simulated were mutation and drift. The first 1,000 historical generations had a constant size of 1,000 per generation and then linearly decreased from 1,000 to 320 in the last hundred historical generations to account for population bottlenecks. The numbers of individuals from each

sex were equal across the historical generations except the last historical generation which included a random sample of 20 males and 300 females (generation 0).

Using the 20 males and 300 females as founder animals, the population size was expanded by simulating 7 generations (genreations 1-7) with the total population size approximately equal to 2,210. Each dam had one or two progenies within each generation with the probability of 0.95 and 0.05, respectively. As with the historical population, the mating was at random without selection and proportion of male progeny was 50%. The replacement rates of sires and dams were 0.6 and 0.2, respectively. Phenotypes with heritability levels of 0.2 and 0.8 were simulated with phenotypic variance of 1.0, where the overall heritability was accounted for by the variance of QTL additive genetic effects assuming no extrta polygenic effect. Allelic effects of QTLs were sampled from a gamma distribution with a shape parameter of 0.4 and a corresponding scale parameter to ensure that the sum of QTLs variances was equal to the predefined QTLs variances. The residual effects were randomly sampled from a Gaussian distribution with a mean of 0 and variance equal to heritability. The overall phenotypic effects were the sum of QTLs effects and residual effects.

Pedigree information was recorded in the recent population from generations 0 to 7. Genotypic data were simulated for individuals (n = 2,210) in generations 1 to 7 coupled with 5,000 or 50,000 biallelic single nucleotide polymorphisms (SNPs) markers evenly distributed across 29 pairs of autosomes with each chromosome length of 100 cM. The number of autosomes and total chromosome length followed those of the bovine genome. Additionally, 290 or 1,015 randomly distributed QTLs were simulated: the former is equivalent to 10 QTLs per chromosome and the latter corresponds to 35 QTLs per chromosome. Markers and QTLs were simulated with a starting allele frequency of 0.5 and a recurrent mutation rate of 2.5×10^{-5} was used to create mutation-drift equilibrium in historical generations. In generation 1,100, markers and QTLs with minor allele frequency greater than 0.05 were

randomly drawn from the segregating loci. Only SNPs but not QTLs were used to infer measures of connectedness and to assess accuracy of prediction.

4.3.2 Management units simulation

The management units were simulated in two steps following Yu et al. [83]: 1) individuals were classified into clusters and 2) clusters were assigned to management units (Figure 4.2). First, 10 individuals were chosen to represent medoids and then 10 distinctive groups were formed by assigning the remaining individuals to the closest medoid using the k-medoid algorithm [64]. The size of 10 distinctive groups ranged from 91 to 590, varying slightly between replications. A dissimilarity matrix was created from the **A** (numerator relationship) matrix by calculating the distance between highest similarity and each similarity coefficient such that the largest similarity coefficient becomes zero. Clustering based on the k-medoid algorithm coupled with the dissimilarity matrix resulted in higher relationship coefficients within a cluster than between clusters.

Two management units were simulated with individuals within clusters assigned to a management unit in six ways. In Scenario 1, a least connected design was simulated by assigning individuals within clusters 1 to 5 into management unit 1 (MU1) and clusters 6 to 10 into management unit 2 (MU2). In scenarios 2 to 6 the degree of genetic link was gradually increased by exchanging 10%, 20%, 30%, 40% and 50% of randomly sampled individuals between MU1 and MU2.

4.3.3 Prediction error variance

Prediction error variance (PEV) can be derived from a linear mixed model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$, $\mathbf{b}$, $\mathbf{g}$ and $\boldsymbol{\epsilon}$ refers to a vector of phenotypes, fixed effects, random additive genetic effects, and residuals, respectively. The incidence matrices $\mathbf{X}$ and $\mathbf{Z}$ connect fixed effects and random additive genetic effects with phenotypes. The joint distribution of random effects is

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{g} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N \left[\begin{pmatrix} \mathbf{Xb} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{ZK}\sigma_g^2\mathbf{Z}' + \mathbf{I}\sigma_\epsilon^2 & \mathbf{ZK}\sigma_g^2 & \mathbf{I}\sigma_\epsilon^2 \\ \mathbf{K}\sigma_g^2\mathbf{Z}' & \mathbf{K}\sigma_g^2 & 0 \\ \mathbf{I}\sigma_\epsilon^2 & 0 & \mathbf{I}\sigma_\epsilon^2 \end{pmatrix} \right],$$

where σ_g^2 is the additive genetic variance, σ_ϵ^2 is the residual variance, and $\mathbf{K}$ represents a relationship matrix, which will be defined in a later section. Following the mixed model equation of Henderson [19]

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{K}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \tag{4.1}$$

where λ is a ratio of variance components which equals to $\frac{\sigma_\epsilon^2}{\sigma_g^2}$. BLUP of $\mathbf{g}$ is given by

$$\widehat{\mathbf{g}} = (\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{K}^{-1}\lambda)^{-1}\mathbf{Z}'\mathbf{M}\mathbf{y},$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is the absorption matrix for fixed effects. Then, the PEV of $\mathbf{g}$ is given by [19]

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

where $\mathbf{C}^{22}$ denotes the lower right quadrant of the inverse of coefficient matrix in equation (4.1).

4.3.4 Genetic connectedness

Two statistics applied in Yu et al. [83] were used to measure connectedness in this study. The first one is the prediction error variance of the differences (PEVD) of EBV between individuals from different management units [20]. A pair-wise comparison between ith and jth individuals is given by the variance of $\hat{g}_i - \hat{g}_j$

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

where ii and jj refer to the diagonal elements of the $\mathbf{C}^{22}$ matrix corresponding to ith and jth individuals, respectively, and ij denotes the off-diagonal elements of $\mathbf{C}^{22}$ matrix. The summary connectedness of PEVD across all pairs of comparisons in a contrast notation is defined as [21]

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

where the sum of elements in a contrast vector $\mathbf{x}$ is zero. For instance, a pair-wise comparison between i'th and j'th management units with $n_{i'}$ and $n_{j'}$ individuals, the contrast vector $\mathbf{x}$ will be set as $1/n_{i'}$, $-1/n_{j'}$ and 0 corresponding to individual belonging to i'th, j'th, and remaining units. The boundary of PEVD is not restricted, with a lower value indicating stronger connectedness. To express connectedness independent from unit of measurement,

PEVD was scaled by additive genetic variance [14, 83].

The generalized coefficient of determination (CD) measures the precision of EBV [21]. Different from PEVD, CD penalizes connectedness measurements if the genetic variability is too small across populations.

$$\begin{aligned}\mathrm{CD}_{ij} &= \frac{var(\mathbf{g}) - var(\mathbf{g}|\widehat{\mathbf{g}})}{var(\mathbf{g})} \\ &= 1 - \frac{var(\mathbf{g}|\widehat{\mathbf{g}})}{var(\mathbf{g})} \\ &= 1 - \lambda \frac{\mathbf{C}^{22}_{ii} + \mathbf{C}^{22}_{jj} - 2\mathbf{C}^{22}_{ij}}{\mathbf{K}_{ii} + \mathbf{K}_{jj} - 2\mathbf{K}_{ij}}.\end{aligned}$$

where CD_{ij} denotes a pair-wise comparison between ith and jth individuals. A summary CD of contrast between any management unit is defined as [67]

$$\begin{aligned}\mathrm{CD}(\mathbf{x}) &= 1 - \frac{var(\mathbf{x}'\mathbf{g}|\widehat{\mathbf{g}})}{var(\mathbf{x}'\mathbf{g})} \\ &= 1 - \lambda \frac{\mathbf{x}'\mathbf{C}^{22}\mathbf{x}}{\mathbf{x}'\mathbf{K}\mathbf{x}},\end{aligned}$$

where $\mathbf{x}$ is the vector of contrast defined earlier. This statistic ranges from 0 to 1 and measures the accuracy of the design. A larger value suggests a stronger estimate of connectedness among management units.

4.3.5 Relationship matrix

Any kind of (semi) positive definite relationship matrices can be used to define $\mathbf{K}$ [85]. We used three types of $\mathbf{K}$ in this study constructed from different sources. The numerator relationship matrix ($\mathbf{K} = \mathbf{A}$) measures the expected additive genetic relationship coefficient between individuals on the basis of pedigree information. The diagonal elements are $1 + F$

, where F represents inbreeding coefficient and off-diagonal elements are equal to twice the kinship coefficients. The construction of the $\mathbf{A}$ matrix was based on tracing all individuals extending over 8 generations to account for historical information and animals from generation 1 to 7 were used for analysis. This matrix expresses relationships as identical by descent (IBD) as it measures the probability of alleles inherited from the same ancestor by tracing pedigree [68].

In contrast, a genomic relationship matrix ($\mathbf{K} = \mathbf{G}$) measures the molecular similarity among individuals. A typical $\mathbf{G}$ matrix is obtained as a function of the gene content matrix ($\mathbf{S}$) including elements of 0, 1, and 2 corresponding to the number of reference alleles. The distribution of jth marker follows the binomial distribution of $s_{\cdot j} \sim \mathrm{B}\left(2p_j, 2p_j(1-p_j)\right)$, where p_j is the allele frequency of jth marker. The $\mathbf{G}$ matrix of VanRaden [69] is obtained as

$$\mathbf{G} = \frac{\mathbf{W}\mathbf{W}'}{m},$$

where $w_{\cdot j}$ is the standardized gene content equal to $\frac{s_{\cdot j} - 2p_j}{\sqrt{2p_j(1-p_j)}}$ and m is the total number of markers.

One item that needs to be addressed when the $\mathbf{A}$ and $\mathbf{G}$ matrices are compared is that they are not on the same scale. For instance, the $\mathbf{A}$ matrix represents relationships among individuals and inbreeding level as deviations from the unrelated base population, conversely the $\mathbf{G}$ matrix expresses those relationships relative to the allele frequencies in the current generation. The following $\mathbf{K} = \mathbf{G}^*$ matrix rescales $\mathbf{G}$ to the same base population as in $\mathbf{A}$ by adjusting the inbreeding coefficient level in $\mathbf{G}$ similar to that of $\mathbf{A}$

$$\mathbf{G}^* = (1 - \bar{F})\mathbf{G} + 2\bar{F}\mathbf{J},$$

where $\bar{F}$ and $\mathbf{J}$ refer to the average inbreeding coefficient of whole population in the $\mathbf{A}$ matrix and the n $\times$ n square matrix filled with 1, respectively [71].

4.3.6 Whole-genome prediction model

The relationship between connectedness and prediction accuracy was investigated with a standard BLUP model

$$\mathbf{y} = 1\boldsymbol{\mu} + \mathbf{g} + \boldsymbol{\epsilon}, \tag{4.2}$$

where $\mathbf{y}$, $\boldsymbol{\mu}$, $\mathbf{g}$ and $\boldsymbol{\epsilon}$ refer to a vector of observed phenotypes, intercept,random additive genetic effects, and residuals, respectively. The model was treated under a Bayesian framework, where $\boldsymbol{\mu}$ was set as a flat prior, with the prior distributions for genetic and residual effects

$$\begin{pmatrix} \mathbf{g} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{K}\sigma_g^2 & 0 \\ 0 & \mathbf{I}\sigma_\epsilon^2 \end{pmatrix} \right],$$

where $\mathbf{K}$ is one of three (semi) positive definite relationship matrices described earlier and $\mathbf{I}$ refers to the identity matrix. The variance components σ_g^2 and σ_ϵ^2 represent variance of additive genetic effects and residual variance, respectively. The scaled inverse χ^2 distribution was assigned to σ_g^2 and σ_ϵ^2 by setting the degrees of freedom (df) equal to 5 and choosing the scale parameter S by equating the mode of scaled inverse χ^2 distribution $\frac{S}{df+2}$ to the quantity of $\frac{R^2 V_y}{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2}$, where R^2 is the expected proportion of phenotypic variance (V_y) explained by the regression and $n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2$ refers to the average sum squares of the genotypes [86]. Here R^2 was set to 0.5 according to Pérez and de los Campos [86].

The prediction accuracy was evaluated by two-fold cross-validation (CV), where the two

management units were treated as the training and testing sets instead of randomly partitioning all individuals into two sets. The variance components were inferred from the data and the predictive ability of the model was calculated as the Pearson correlation between predicted genetic values and true genetic values in the testing set. Throughout this study, the BGLR R package was used to fit equation (4.2). A Gibbs sampler was run for 10,000 iterations, where the first 2,000 samples were discarded as burn-in. A total of 8,000 samples coupled with a thinning rate of 5 were used to infer posterior means.

4.3.7 Criterion for connectedness measures

The challenge with discussing connectedness is that there is no clear standard or benchmark for true connectedness. Although zero connectedness may be an indicator of possible bias, this issue has been discussed since Foulley et al. [9]. In this respect, Kuehn et al. [14] proposed threshold values for moderate and strong levels of connectedness based on the relationship between prediction error correlation and model-based mean squared error. In this study, we provide a guideline for connectedness measures in terms of whole-genome prediction by performing CV. Note that prediction accuracy may simply increase as PEVD continues to decrease no matter how individuals across management units become genetically alike. On the other hand, measures of CD start to decrease as in Yu et al. [83] when across management units include individuals that are too genetically similar. CD is suited for deriving a criterion because there is no point in enhancing prediction accuracy by simply reducing relatedness variability. Therefore, we explored the approximate threshold of CD that yields a reasonable prediction accuracy while maintaining genetic diversity in a population [21, 67].

4.4 Results

Figure 4.3 displays relationships between two management units with 5,000 markers used to compute three relationship matrices (**A**, **G**, and $\mathbf{G}^*$) according to six simulated management unit scenarios. For each scenario, average relationships were the highest for **A** and the smallest for **G**, and $\mathbf{G}^*$ produced relationships somewhere between **A** and **G**. Relationships increased when more individuals were exchanged between the two units. This increasing relationship pattern was observed regardless of relationship matrices used. A similar tendency was shown when the number of markers was equal to 50,000 (result not shown).

4.4.1 Prediction error variance of the difference

The relationships between measures of connectedness and prediction accuracies obtained from the Bayesian BLUP model are shown in Figures 4.4 and 4.5. The prediction accuracies in Figures 4.4 and 4.5 are identical as they are based on the same simulations. Figure 4.4 depicts connectedness measured as PEVD of contrast with smaller values inferring increased connectedness. Generally, increased connectedness measures and prediction accuracies were observed as more individuals from the same clusters were shared between management units, regardless of h^2 levels, type of kernel matrices, the number of QTLs, and marker density. Similarly, standard errors of estimates over ten replicates ranged from 0.008 to 0.068 for prediction accuracy, and from 0.001 to 0.002 for PEVD, regardless of h^2 levels, type of kernel matrices, the number of QTLs, and marker density. In Figure 4.4A with 290 QTLs and 5,000 markers, the **G** and $\mathbf{G}^*$ matrices delivered similar or stronger connectedness measures and higher prediction accuracies than those of the **A** matrix. The results from $\mathbf{G}^*$ strongly resembled those of **G** in terms of measures of connectedness and prediction accuracies. When marker density increased to 50,000, with the same number of QTLs, slightly improved pre-

diction accuracies and increased estimates of connectedness were observed (Figure 4.4B). Stronger connectedness and higher prediction accuracy were shown with $\mathbf{G}$ and $\mathbf{G}^*$ than $\mathbf{A}$. The pattern in Figure 4.4C with 1,015 QTLs and 5,000 markers resembled that of Figure 4.4A, however, we observed marginally decreased genomic prediction accuracies. Figure 4.4D with 1,015 QTLs and 50,000 markers presented the clearest pattern: the $\mathbf{G}$ and $\mathbf{G}^*$ matrices consistently produced stronger estimates of connectedness and higher prediction accuracies than those of the $\mathbf{A}$ regardless of simulation scenarios and h^2 levels.

4.4.2 Coefficient of determination

The change of prediction accuracies with the increasing proportion of linked individuals quantified with CD of contrast is shown in Figure 4.5, where larger CD values suggest stronger connectedness. The standard errors of estimates for CD through ten replicates varied from 0.004 to 0.057, regardless of h^2 levels, type of kernel matrices, the number of QTLs, and marker density. In general, the prediction accuracy improved when more individuals from the same clusters were assigned across units. Within each scenario, the estimates of CD increased up to Scenario 3 and decreased at Scenario 4 because CD penalized connectedness measures for reduced genetic variability. This corresponded to 20% exchange rate.

In Figure 4.5A with 290 QTLs and 5,000 markers, similar or stronger connectedness and higher prediction accuracies were observed by the $\mathbf{G}$ matrix than those using $\mathbf{A}$ for all scenarios. An analogous tendency was identified in Figure 4.5C with 1,015 QTLs and 5,000 markers, except that marginal reduction of genomic prediction accuracies were observed. With 290 QTLs and an increased number of markers (50,000), both genomic prediction accuracies and estimates of connectedness increased slightly (Figure 4.5B). Overall, $\mathbf{G}$ and

$\mathbf{G}^*$ presented stronger estimates of connectedness and higher prediction accuracies than those of $\mathbf{A}$. Clearer differences were observed when increasing the number of QTLs to 1,015 (Figure 4.5D). The $\mathbf{G}$ matrix clearly yielded higher estimates of connectedness and higher prediction accuracies as compared to $\mathbf{A}$. The performances of $\mathbf{G}^*$ were very similar to those of $\mathbf{G}$ in CD across all cases.

4.5 Discussion

The concept of connectedness dates back to estimability in experimental design in the sense of all-or-none connectedness [87, 88]. A dataset can be seen as connected if merging cells in a cross-table is possible such that all filled cells are connected [89]. It was later extended to a random effect model or BLUP genetic evaluation known as reference sire progeny testing schemes by Foulley et al. [9, 12] and Miraei Ashtiani and James [90]. The central idea is when sires from one management unit are compared against sires in another unit, at least one sire should be tested in both units. Such common sires are known as link sires or reference sires. These authors investigated the efficient strategy of reference sire use to minimize PEVD between EBV by identifying the optimal number of progeny. Since then connectedness based on pedigree information has taken center stage in both theoretical development and real data applications [e.g., 14, 21, 91]. In addition, non-PEV-based genetic connectedness metrics have been developed [e.g., 78]. Connectedness is often used as an indicator of the robustness of genetic evaluation comparisons, where a higher level of connectedness suggests more reliable comparison of EBV across units. Past studies found that BLUP evaluations correctly yielded the likely ranking of individuals distributed across units when connectedness was present. While research in pedigree-based connectedness is still critical, as shown in Yu et al. [83] and in the current study, availability of genomic information now offers an

opportunity to revisit a number of critical questions related to connectedness, such as how prediction accuracy is influenced given the level of connectedness between management units.

The extent of connectedness level boils down to the ability of **K** to capture relationships among individuals. Connectedness increases with stronger across unit genetic relationship and it decreases with stronger within unit relationship [20]. Advantages of genomic over pedigree relationships are as follows: 1) genomic measures relatedness arising from more distant ancestors than those included in a pedigree and 2) genomic captures the variation in realized kinship arising from the stochastic effects of Mendelian sampling and recombination. We tested three types of **K** to capture the relationship among individuals in this study. The two matrices **A** and **G** mainly differ in 1) the distinction between IBD and IBS and 2) the relationships are relative to the baseline population versus current population. The $\mathbf{G}^*$ relationship matrix helps to put **A** and **G** on a similar scale. Although those factors contributed to the improved quality of genetic evaluation design with the increased proportion of connecting individuals as shown in Yu et al. [83], the relationship between connectedness level and CV derived prediction accuracy has been yet-to-be answered. The present study aimed to bridge this gap by applying PEVD and CD of contrasts to simulated phenotypes, pedigrees, genomics, and management units. Note that the magnitude of the differences in results may be observed when applied to real data compared to the simulation results shown in this study.

4.5.1 Relationship between connectedness and prediction accuracy

We used contrasts of PEVD and CD to investigate the relationship between connectedness and prediction accuracy. We found prediction accuracy improved with increased capturing

of connectedness between units. This suggests that increase in the accuracy of the EBV comparison is positively associated with increase in accuracy of CV-based prediction. In general, genomic prediction accuracy improved as more markers were used to infer a genomic relationship matrix and as more QTLs contributed to the genetic variation given plenty of markers. These can be attributed to the fact that 1) the greater the number of markers, the better capturing of QTL relationships among individuals [92] and 2) genomic best linear unbiased prediction performs better when the number of QTLs is large, because of its infinitesimal model assumption [93]. This result may change when an alternative whole-genome prediction model is used instead of genomic best linear unbiased prediction. For instance, a BayesB type of model performs well when the number of QTLs is small [93]. Measures of connectedness increased as more markers were used to characterize connectedness. When more markers were used, genomic information captures more variation in relationships which results in increased measures of connectedness.

Across six management unit scenarios, the extent of connectedness measured by PEVD and prediction accuracy from BLUP were higher as the proportion of individuals exchanged between the two units increased. The measurement of PEVD decreases when the number of markers increase regardless of QTL numbers and h^2 levels. This was not always the case in CD because this statistic penalizes connectedness estimates when the amount of genetic variability across units was small.

The $\mathbf{G}$ and $\mathbf{G}^*$ matrices clearly outperformed that of $\mathbf{A}$ in prediction and also produced increased measures of connectedness (Figures 4.4 and 4.5). Interestingly, although the average relationship of individuals across management units computed from the $\mathbf{G}^*$ matrix was more similar with that of $\mathbf{A}$ than $\mathbf{G}$ (Figures 4.3), the results of connectedness estimates and prediction accuracies obtained from the $\mathbf{G}^*$ matrix were more similar with those of $\mathbf{G}$ (Figures 4.4 and 4.5). This is most likely because of the similar variation in relationships

across management units captured by $\mathbf{G}$ and $\mathbf{G}^*$, which play an important role in measures of connectedness and prediction accuracies. The effect of scaling $\mathbf{G}$ to be more similar to $\mathbf{A}$ was minimal for PEVD and CD as $\mathbf{G}^*$ produced increased measure of connectedness compared to that of $\mathbf{A}$. This is in agreement with Yu et al. [83] where they found that genome-based connectedness consistently increased estimates of connectedness in most cases regardless of rescaling $\mathbf{G}$ to the level of $\mathbf{A}$.

In addition, we observed marginally decreased genomic prediction accuracies when the number of QTLs was increased while the number of SNPs remained constant (Figures 4.4A vs. 4.4C and 4.5A vs 4.5C). This is because the number of parameters we need to accurately predict increased and a sufficient number of markers is required to establish a sufficient level of LD to capture QTL signals. With more QTL, more markers are needed for them to contribute to or enhance prediction accuracy. This observation can be also supported theoretically from interactive deterministic genomic prediction accuracy simulators [94].

4.5.2 What is the sufficient level of connectedness?

The extent to which a design is genetically connected or not has been the subject of discussion in the literature [e.g., 95, 96]. These authors proposed statistical approaches to determine the presence or absence of connectedness. A related question is to find a desired or sufficient level of connectedness based on connectedness metrics as in Kuehn et al. [14]. Here CD statistic offers an important insight because it accounts for the reduction of connectedness due to reduced genetic variability between individuals under comparison. This pattern was also observed by using both pedigree and genome-based CD connectedness in Yu et al. [83]. From the perspective of designing a breeding program, increasing connectedness simply by making individuals genetically similar to each other should be avoided [21]. Thus, the use

of CD allows us to identify an upper limit of sufficient CD value that gives a reasonable prediction accuracy while maintaining the variability of relatedness. The CD began to fall around 20% exchange rate and the threshold CD value was in the range of 0.7 to 0.9 across simulation scenarios. When the measures of CD exceeded this threshold, prediction accuracy continued to improve in a mild degree or stayed the same, whereas connectedness estimates started to decrease. Although this cutoff value slightly varies among different scenarios [83], the CD metric can be used to optimize selective genotyping and phenotyping along the lines of Rincent et al. [57] and Isidro et al. [58]. In contrast, when connectedness was determined with PEVD, prediction accuracy and connectedness both continued to increase when shifting more individuals across management units thereby increasing genetic similarity. Such is clearly not a desired property in designing a breeding program.

4.6 Conclusions

In general, connectedness measures and prediction accuracies increased as more individuals from the same clusters were shared across management units. We found prediction accuracy improved with increased capturing of connectedness across units suggesting that increase in the accuracy of the EBV comparison is positively associated with increase in accuracy of CV-based prediction. This was entirely true for PEVD and partly so for CD. The impact of genomics was more marked compared to pedigree when a sufficient number of markers was present to capture QTLs. While there is a need to establish increased levels of connectedness, simply increasing connectedness results in rapid decrease of relatedness variability which may not be desired in a breeding program. Use of CD allows us to find a connectedness level that gives a reasonable prediction accuracy while maintaining genetic diversity in a population.

4.7 Figures

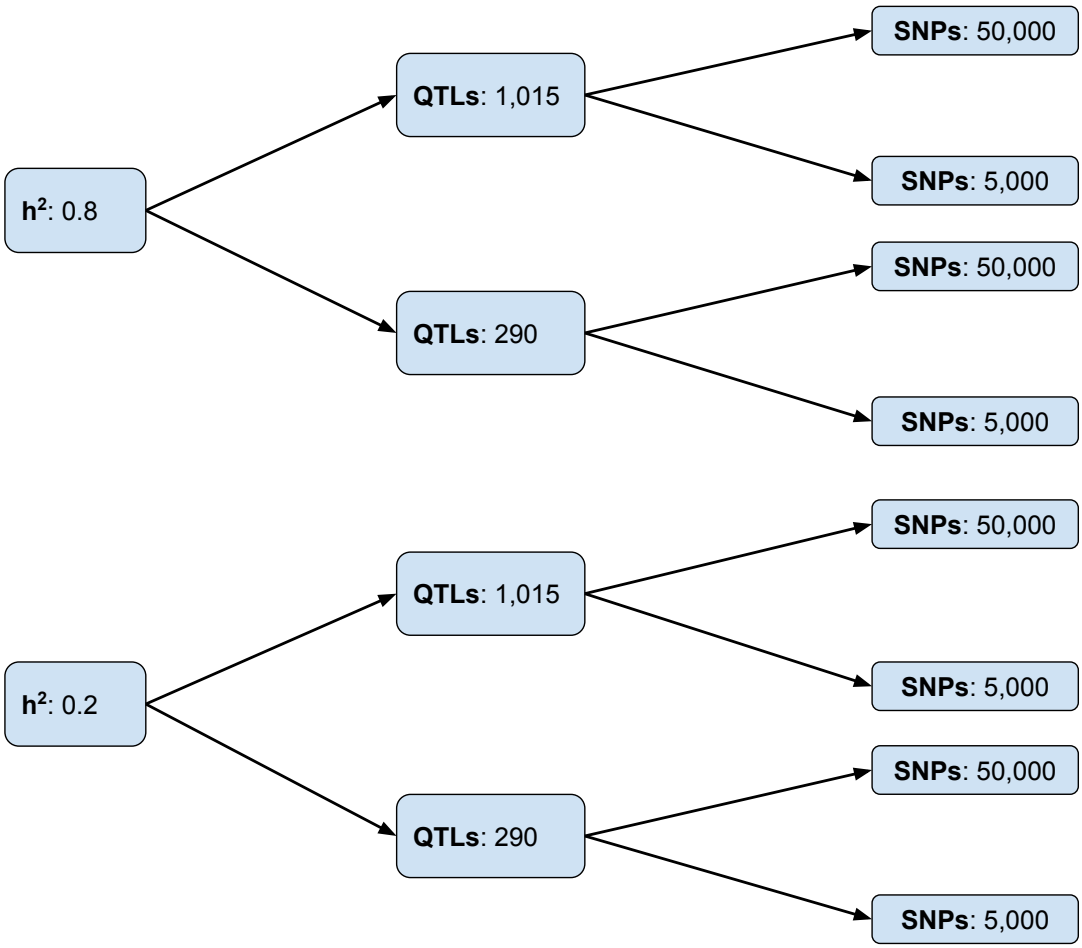

Figure 4.1: Genomic data simulation parameters. SNPs, QTLs and h^2 represent total single-nucleotide polymorphisms, quantitative trait loci, and trait heritability, respectively. Simulations were carried out across two different h^2 (0.8 and 0.2), two different numbers of QTLs (1,015 and 290) and two different SNP densities (50,000 and 5,000).

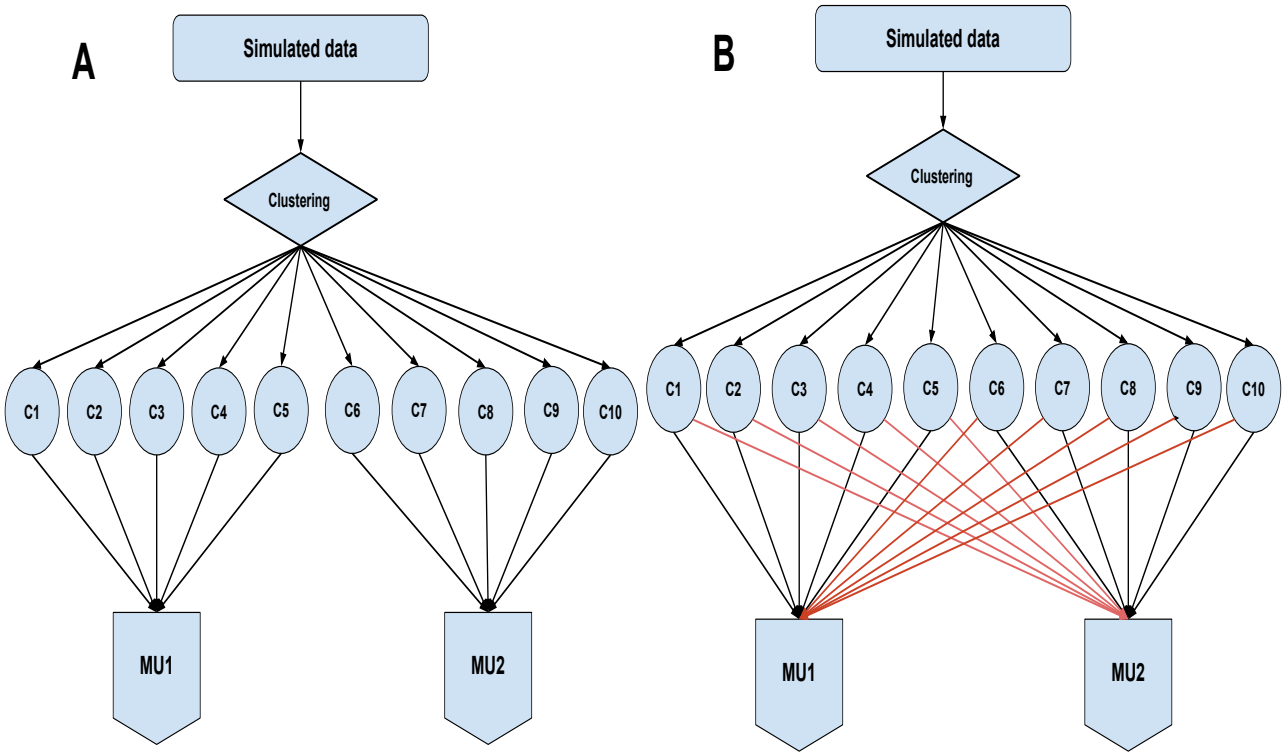

Figure 4.2: Management unit (MU) simulation scenarios. A: Scenario 1 (least connected design). Individuals within clusters 1 to 5 were assigned to MU1 and clusters 6 to 10 were assigned to MU2. B: Scenarios 2 to 6 (partially connected to connected). The degree of connectedness was gradually increased by exchanging 10% (Scenario 2), 20% (Scenario 3), 30% (Scenario 4), 40% (Scenario 5) and 50% (Scenario 6) of randomly sampled individuals between MU1 and MU2. Scenario 6 corresponds to the connected design.

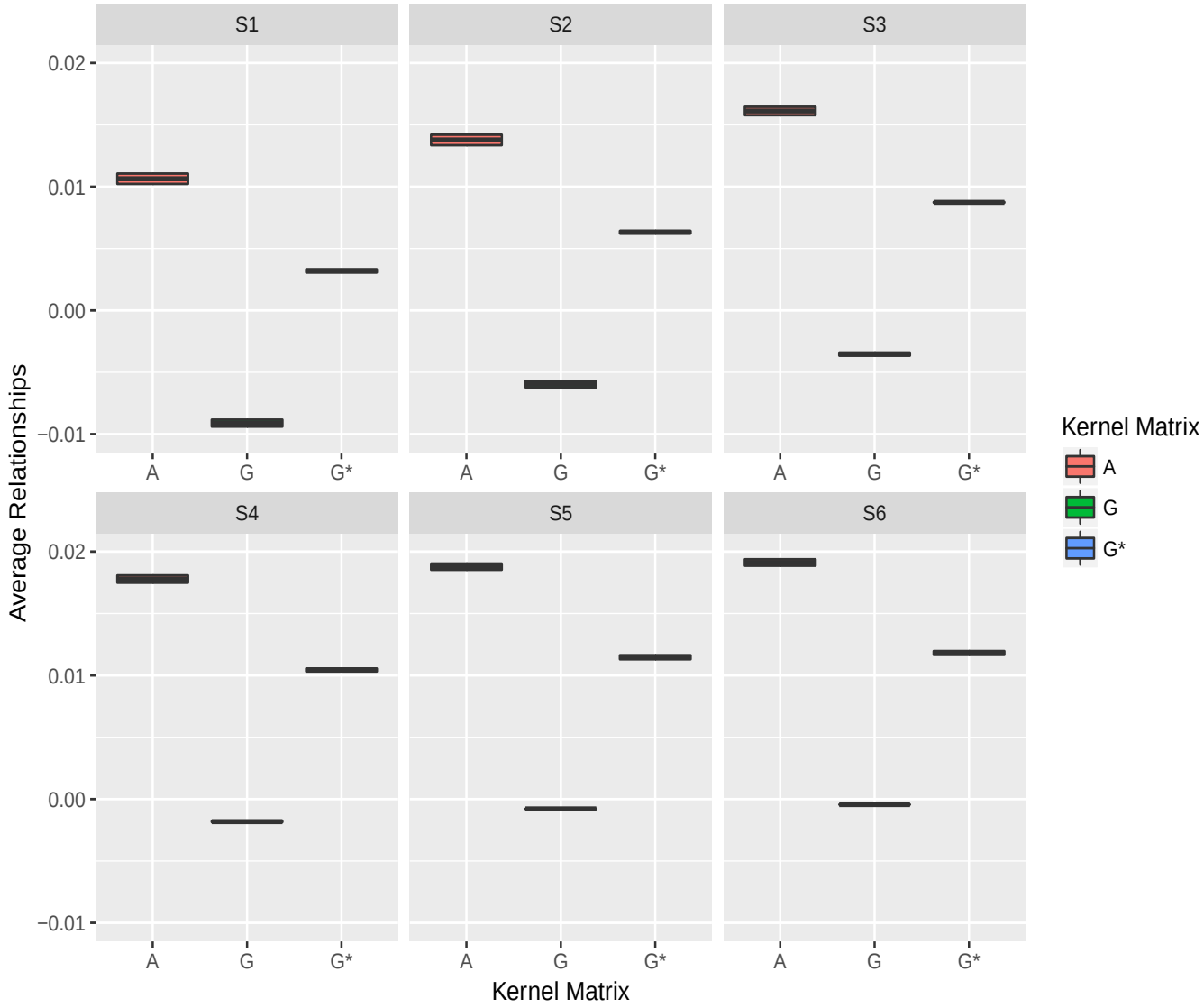

Figure 4.3: Average relationship coefficients across management units with 5,000 markers over two heritability levels and two different numbers of quantitative trait loci. S1 to S6 denotes management unit simulation scenario 1, 2, 3, 4, 5 and 6, respectively. The magnitude of connectedness level steadily increased from S1 to S6. We compared pedigree-based $\mathbf{A}$, genome-based $\mathbf{G}$, and rescaled genome-based $\mathbf{G}^*$ relationship kernel matrices.

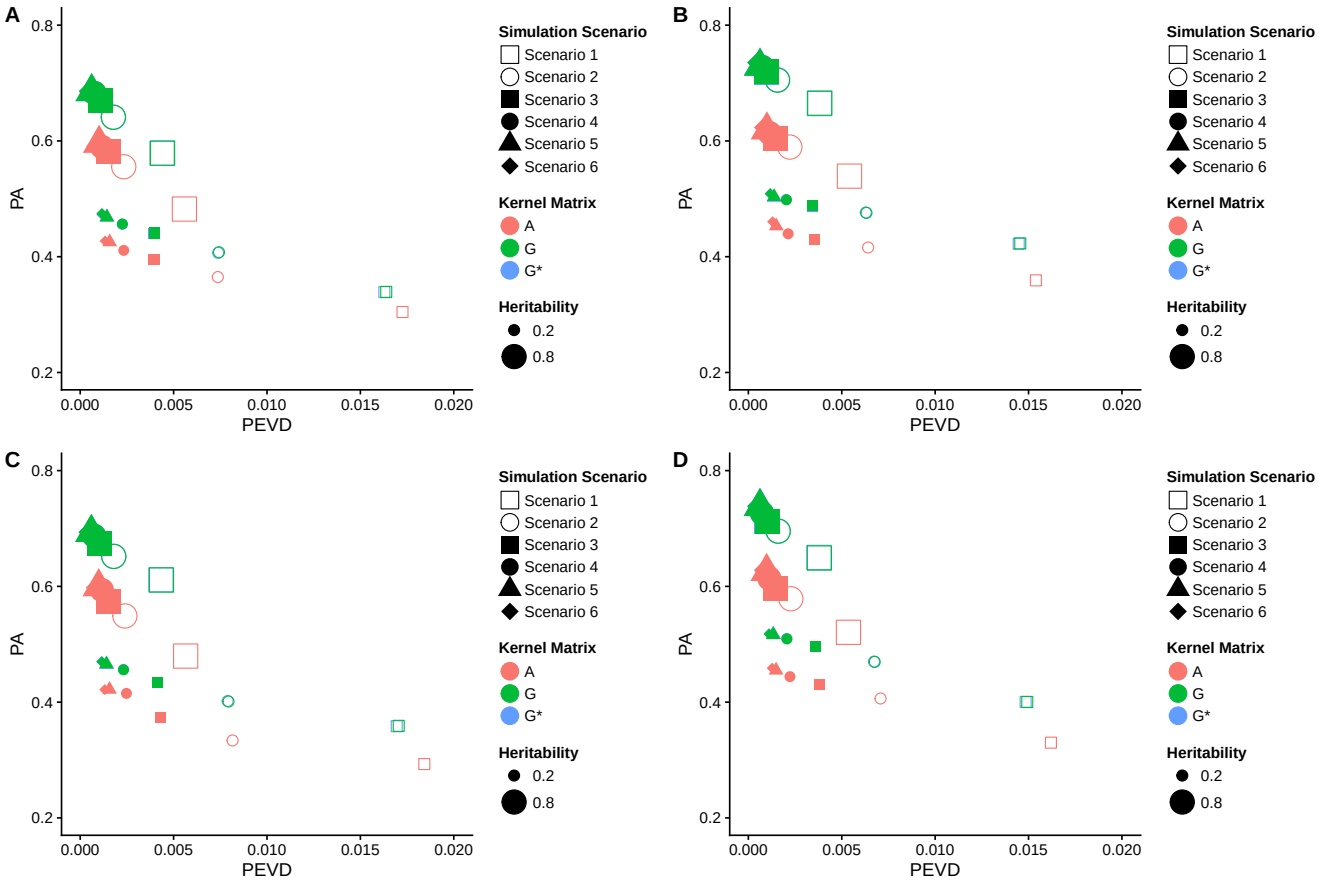

Figure 4.4: Relationship between connectedness and prediction accuracy. PEVD and PA denote prediction error variance of the differences and prediction accuracy, respectively. PA was defined as the correlation between phenotypes and estimated breeding values $cor(\mathbf{g}, \hat{\mathbf{g}})$. Connectedness of pedigree-based $\mathbf{A}$, genome-based $\mathbf{G}$, and rescaled genome-based $\mathbf{G}^*$ within 6 management units simulation scenarios across 2 heritabilties were compared with their prediction accuracies in each graph. A: 290 QTLs and 5,000 markers. B: 290 QTLs and 50,000 markers. C: 1,015 QTLs and 5,000 markers. D: 1,015 QTLs and 50,000 markers.

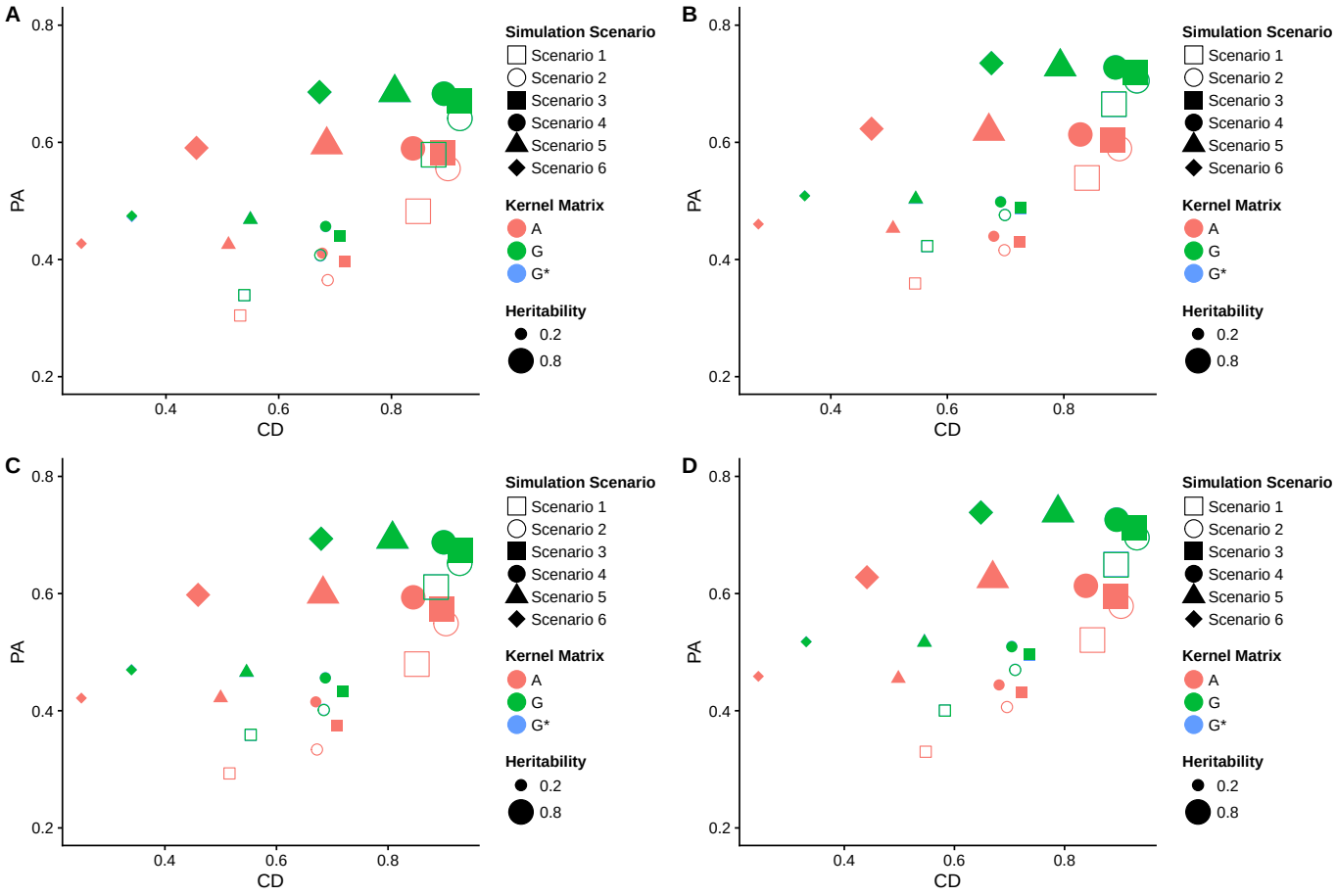

Figure 4.5: Relationship between connectedness and prediction accuracy. CD and PA denote coefficient of determination and prediction accuracy, respectively. PA was defined as the correlation between phenotypes and estimated breeding values $cor(\mathbf{g}, \hat{\mathbf{g}})$. Connectedness of pedigree-based $\mathbf{A}$, genome-based $\mathbf{G}$, and and rescaled genome-based $\mathbf{G}^*$ within 6 management units simulation scenarios across 2 heritabilties were compared with their prediction accuracies in each graph. A: 290 QTLs and 5,000 markers. B: 290 QTLs and 50,000 markers. C: 1,015 QTLs and 5,000 markers. D: 1,015 QTLs and 50,000 markers.

Chapter 5

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

5.1 Abstract

Background: Genetic connectedness is a critical component of genetic evaluation as it assesses the comparability of predicted genetic values across management 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.

5.2 Introduction

Genetic connectedness quantifies the extent to which estimated breeding values can be fairly compared across management units or contemporary groups [12, 97]. 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 [98]. 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 [99]. The concept of genetic connectedness was later extended to measure the connectedness level between reference and validation sets in whole-genome prediction. In general, it was observed that increased connectedness led to increased prediction accuracy of genetic values evaluated by a cross-validation [100]. Comparability of total genetic values across units by accounting for additive as well as non-additive genetic effects has also been investigated [101].

Despite the importance of connectedness, there is no user-friendly software tool available that offers computation of a comprehensive list of connectedness statistics. Therefore, we developed a genetic connectedness analysis R package, GCA, which measures the connectedness between individuals across units using pedigree and 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.

5.3 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 at the unit level as the average PEV of all pairwise differences between individuals across units, average PEV within and 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.

5.3.1 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

[19]. The $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices associating systematic effects and genetic values to observations, respectively. The joint distribution of random effects is as given below.

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N \left[\begin{pmatrix} \mathbf{Xb} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{ZK}\sigma_u^2\mathbf{Z}' + \mathbf{I}\sigma_\epsilon^2 & \mathbf{ZK}\sigma_u^2 & \mathbf{I}\sigma_\epsilon^2 \\ \mathbf{KZ}'\sigma_u^2 & \mathbf{K}\sigma_u^2 & 0 \\ \mathbf{I}\sigma_\epsilon^2 & 0 & \mathbf{I}\sigma_\epsilon^2 \end{pmatrix} \right],$$

where $\mathbf{K}$ is a relationship matrix, σ_u^2 is the additive genetic variance, and σ_e^2 is the residual variance. The corresponding MME is as given below.

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{K}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix},$$

where $\lambda = \frac{\sigma_\epsilon^2}{\sigma_u^2}$ is the ratio of variance components. The inverse of the MME coefficient matrix derived from this model is as given below.

$$\begin{aligned} \mathbf{C}^{-1} &= \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{K}^{-1}\lambda \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix}. \end{aligned}$$

Then the PEV of $\mathbf{u}$ is derived as shown in Henderson [19].

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

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is the absorption (projection) matrix for fixed effects and $\mathbf{C}^{22}$ is the lower right quadrant of the inverse of coefficient matrix. Note that 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 management unit or contemporary group effects. Kennedy and Trus [102] argued that mean PEV over unit (PEV_{Mean}) defined as the average of PEV between individuals within the same unit can be approximated by VE = $\text{Var}(\hat{b})$, that is

$$\begin{aligned} \text{VE0} &= \text{Var}(\hat{b}) \\ &= [\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \mathbf{K}^{-1}\lambda)^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\sigma_{\epsilon}^{2} \\ &\approx \text{PEV}_{\text{Mean}} \end{aligned} \tag{5.1}$$

Holmes et al. [23] pointed out that the agreement between PEV_{Mean} and VE0 depends on a number of fixed effects other than the management group fitted in the model. They proposed exact ways to derive PEV_{Mean} as a function of VE and suggested addition of a few correction factors. When unit effect is the only fixed effect included in the model, the exact PEV_{Mean} can be obtained as given below.

$$\text{VE1} = \text{PEV}_{\text{Mean}} = \text{Var}(\hat{b}) - \sigma_{\epsilon}^{2}(\mathbf{X}'\mathbf{X})^{-1}, \tag{5.2}$$

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 $\sigma^2_\epsilon(\mathbf{X}'\mathbf{X})^{-1}$ corrects the number of records within units. Accounting for additional fixed effects beyond unit effect when computing PEV_{Mean} is given by the following equation.

$$\begin{aligned} \text{VE2} &= \text{PEV}_{\text{Mean}} \qquad (5.3) \\ &= \text{Var}(\hat{b}_1) - \sigma^2_\epsilon(\mathbf{X_1}'\mathbf{X_1})^{-1} \\ &+ (\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{X_2}\text{Var}(\hat{b}_2)\mathbf{X_2}'\mathbf{X_1}(\mathbf{X_1}'\mathbf{X_1})^{-1} \\ &+ (\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{X_2}\text{Cov}(\hat{b}_2, \hat{b}_1) \\ &+ \text{Cov}(\hat{b}_1, \hat{b}_2)\mathbf{X_2}'\mathbf{X_1}(\mathbf{X_1}'\mathbf{X_1})^{-1}, \qquad (5.4) \end{aligned}$$

where $\mathbf{X_1}$ and $\mathbf{X_2}$ represent incidence matrices for units and other fixed effects, respectively, and $\hat{b}_1$ and $\hat{b}_2$ refer to the estimates of unit effects and other fixed effects, respectively [23]. This equation is suitable for cases in which there are two or more fixed effects fitted in the model.

5.3.2 Connectedness metrics

Below we describe connectedness metrics implemented in the GCA package. These metrics are the function of PEV or VE described earlier (Figure 5.1).

Prediction error variance of difference (PEVD)

A PEVD metric measures the prediction error variance difference of breeding values between individuals from different units [102]. The PEVD between two individuals i and j is expressed as shown below.

$$\begin{aligned}\text{PEVD}(\hat{u}_i - \hat{u}_j) &= [\text{PEV}(\hat{u}_i) + \text{PEV}(\hat{u}_j) - 2\text{PEC}(\hat{u}_i, \hat{u}_j)] \\ &= (\mathbf{C}^{22}_{ii} - \mathbf{C}^{22}_{ij} - \mathbf{C}^{22}_{ji} + \mathbf{C}^{22}_{jj})\sigma^2_\epsilon \\ &= (\mathbf{C}^{22}_{ii} + \mathbf{C}^{22}_{jj} - 2\mathbf{C}^{22}_{ij})\sigma^2_\epsilon, \end{aligned} \tag{5.5}$$

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

Individual average PEVD: When pairwise PEVD is first computed at the individual level using equation (5.5), these estimates need to be aggregated and summarized at the unit level. A calculation of summary PEVD can be traced back to Kennedy and Trus [102] as the average of PEVD between individuals across two units.

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

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 5.2A.

Group average PEVD: The second summary method applies equation (5.5) after calculating PEV_{Mean} of i'th and j'th units and mean prediction error covariance (PEC_{Mean}) between i'th and j'th units.

$$\text{PEVD}_{i'j'} = \overline{\text{PEV}}_{i'i'} + \overline{\text{PEV}}_{j'j'} - 2\overline{\text{PEC}}_{i'j'}, \tag{5.6}$$

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

in Figure 5.2B.

Contrast PEVD: The third summary method is PEVD of contrast between a pair of units.

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

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 5.2C.

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 $\mathbf{u}$ [21]. A notable difference between CD and PEVD is that CD penalizes connectedness measurements when across units include individuals that are genetically too similar [99, 100]. A pairwise CD between individuals i and j is given by the following equation.

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

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 [103].

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

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

A flow diagram of individual average CD is shown in Figure 5.3A.

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.

$$\begin{aligned}
\mathrm{CD}_{i'j'} &= 1 - \lambda \cdot \frac{\overline{\mathbf{C^{22}}}_{i'i'} + \overline{\mathbf{C^{22}}}_{j'j'} - 2\overline{\mathbf{C^{22}}}_{i'j'}}{(\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})} \\
&= 1 - \frac{\sigma_e^2 \cdot \overline{\mathbf{C^{22}}}_{i'i'} + \overline{\mathbf{C^{22}}}_{j'j'} - 2\overline{\mathbf{C^{22}}}_{i'j'}}{\sigma_u^2 \cdot (\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})} \\
&= 1 - \frac{\overline{\mathrm{PEV}}_{i'i'} + \overline{\mathrm{PEV}}_{j'j'} - 2\overline{\mathrm{PEC}}_{i'j'}}{\sigma_u^2 \cdot (\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})} \\
&= 1 - \frac{\mathrm{PEVD}_{i'j'}}{\sigma_u^2 \cdot (\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})}.
\end{aligned} \tag{5.7}$$

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 5.3B.

Contrast CD: A contrast of CD between any pair of units is given by [103]

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

A flow diagram showing the computational procedure is shown in Figure 5.3C.

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 [22].

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

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.

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

This summary method for r statistic was used in Yu et al. [99] and calculation steps are shown in Figure 5.4A.

Group average r: This is known as flock connectedness in the literature, which calculates the ratio of PEV_{Mean} and PEC_{Mean} [98]. This group average connectedness for r between

two units i' and j' is given by the following equation.

$$\begin{aligned} \mathrm{r}_{i'j'} &= \frac{\overline{\mathrm{PEC}}_{i'j'}}{\sqrt{\overline{\mathrm{PEV}}_{i'i'} \cdot \overline{\mathrm{PEV}}_{j'j'}}} \\ &= \frac{1/n_{i'} \sum \mathrm{PEC}_{i'j'} 1/n_{j'}}{\sqrt{(1/n_{i'})^2 \sum \mathrm{PEV}_{i'i'} \cdot (1/n_{j'})^2 \sum \mathrm{PEV}_{j'j'}}} \\ &= \frac{\sum \mathrm{PEC}_{i'j'}}{\sqrt{\sum \mathrm{PEV}_{i'i'} \cdot \sum \mathrm{PEV}_{j'j'}}}. \end{aligned} \tag{5.8}$$

A graphical derivation is presented in Figure 5.4B.

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

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

A flow diagram illustrating a computational procedure is shown in Figure 5.4C.

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. 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 [23]. Note that all VE-derived approaches can be classified based on the number of fixed effects to be corrected including no correction (0), correction for one fixed effect (1), and correction for two or more fixed effects (2) [23]. Below we discuss connectedness metrics that are derived from VED.

VED0: Using the summary method group average, the VED0 statistic [102] estimates PEVD alike connectedness with VE rather than PEV_{Mean}. We can obtain VED0 between two units i' and j' by replacing PEV_{Mean} in equation (5.6) with VE0 defined in equation (5.1).

$$\text{VED0}_{i'j'} = \text{VE0}_{i'i'} + \text{VE0}_{j'j'} - 2\text{VE0}_{i'j'}, \tag{5.9}$$

VED1: A VED statistic that corrects for the presence of unit effect is obtained by replacing PEV_{Mean} in equation (5.6) with VE1 defined in equation (5.2). This corrects for the number of individuals in the units.

$$\text{VED1}_{i'j'} = \text{VE1}_{i'i'} + \text{VE1}_{j'j'} - 2\text{VE1}_{i'j'}, \tag{5.10}$$

VED2: Similarly, VED statistic based on VE2 is obtained by replacing PEV_{Mean} in equation (5.6) with VE2 defined in equation (5.3). This formula accounts for fixed effects other than unit effect.

$$\text{VED2}_{i'j'} = \text{VE2}_{i'i'} + \text{VE2}_{j'j'} - 2\text{VE2}_{i'j'}, \tag{5.11}$$

Coefficient of determination of VED

CDVED0: A CDVED0 statistic, which is a CD statistic based on VE0, is defined by replacing PEV_{Mean} in equation (5.7) with VE0. A pairwise CDVED0 between two units i' and j' is given by the following equation.

$$\text{CDVED0}_{i'j'} = 1 - \frac{\text{VE0}_{i'i'} + \text{VE0}_{j'j'} - 2\text{VE0}_{i'j'}}{\sigma_u^2 \cdot (\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})}.$$

CDVED1: CDVED1 is obtained by replacing PEV_{Mean} in equation (5.7) with VE1.

$$\text{CDVED1}_{i'j'} = 1 - \frac{\text{VE1}_{i'i'} + \text{VE1}_{j'j'} - 2\text{VE1}_{i'j'}}{\sigma_u^2 \cdot (\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})},$$

CDVED2: Similarly, CDVED2 is obtained by replacing PEV_{Mean} in equation (5.7) with VE2.

$$\text{CDVED2}_{i'j'} = 1 - \frac{\text{VE2}_{i'i'} + \text{VE2}_{j'j'} - 2\text{VE2}_{i'j'}}{\sigma_u^2 \cdot (\overline{\mathbf{K}}_{i'i'} + \overline{\mathbf{K}}_{j'j'} - 2\overline{\mathbf{K}}_{i'j'})},$$

Connectedness rating (CR)

CR0: A CR statistic first proposed by Mathur et al. [104] is similar to equation (5.8). However, it uses variances and covariances of estimated unit effects. Specifically, we replace PEV_{Mean} with VE0, and CR0 between two units i' and j' is given by the following equation.

$$\text{CR0}_{i'j'} = \frac{\text{VE0}_{i'j'}}{\sqrt{\text{VE0}_{i'i'} \cdot \text{VE0}_{j'j'}}}.$$

CR1: A CR1 statistic is obtained by replacing PEV_{Mean} in equation (5.8) with VE1.

$$\text{CR1}_{i'j'} = \frac{\text{VE1}_{i'j'}}{\sqrt{\text{VE1}_{i'i'} \cdot \text{VE1}_{j'j'}}},$$

CR2: In the same manner, a CR2 statistic is obtained by replacing PEV_{Mean} in equation (5.8) with VE2.

$$\text{CR2}_{i'j'} = \frac{\text{VE2}_{i'j'}}{\sqrt{\text{VE2}_{i'i'} \cdot \text{VE2}_{j'j'}}},$$

5.4 Software Description

5.4.1 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 [105]. 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 [106]. The initial versions of the algorithms and the R code were used in previous studies [99, 100, 101] 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.

5.4.2 Installing the GCA Package

The current version of the GCA R package is available at GitHub (`https://github.com/HaipengU/GCA`). The package can be installed using the devtools R package [107] and loaded into the R environment.

Box 1: Installing the GCA Package

```
install.packages("devtools")
library(devtools)
install_github('HaipengU/GCA')
library(GCA)
```

5.4.3 Simulated data

We simulated a cattle data set using QMSim software [84] to illustrate the usage of GCA package. This data set is included in the 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 [108] coupled with the dissimilarity matrix derived from a numerator relationship matrix as shown in previous studies [99, 100, 101]. The data set was stored as an R object in the package.

Box 2: Loading the data

```
data(package = 'GCA')$results[, "Item"] # list all data files in the GCA package
data(GCcattle) # load the data
dim(cattle.pheno) # phenotype and fixed effects
dim(cattle.W) # marker matrix
```

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.

5.4.4 Application of the GCA Package

Below we show the usage of the main function gca followed by some specific examples using CD. Box 3 lists all input arguments for the gca function.

- Kmatrix Genetic relationship matrix constructed from either pedigree or genomics.
- Xmatrix: Fixed effects incidence matrix excluding intercept. The first column of the Xmatrix should start with unit effects followed by other fixed effects if applicable.
- sigma2a and sigma2e: Estimates of additive genetic and residual variances, respectively.
- MUScenario: A vector of fixed factor units.
- statistic: Choice of connectedness statistic. Available options include
    1. PEV-derived functions: PEVD_IdAve, PEVD_GrpAve, PEVD_contrast, CD_IdAve, CD_GrpAve, CD_contrast, r_IdAve, r_GrpAve, and r_contrast
    2. VE-derived functions: VED0, VED1, VED2, CDVED0, CDVED1, CDVED2, CR0, CR1, and CR2.
- NumofMU: Return either pairwise unit connectedness (Pairwise) or overall connectedness across all units (Overall).
- Uidx: An integer indicating the last column number of units in the Xmatrix. This Uidx is required for VED2, CDVED2, and CR2 statistics. The default is NULL.
- scale (logical): Should sigma2a be used to scale statistic (i.e., PEVD_IdAve, PEVD_GrpAve, PEVD_contrast, VED0, VED1, and VED2) so that connectedness is independent of measurement unit? The default is TRUE.
- diag (logical): Should the diagonal elements of the PEV matrix (i.e., PEVD_GrpAve, CD_GrpAve, and r_GrpAve) or the K matrix (CDVED0, CDVED1, and CDVED2) be included? The default is TRUE.

Box 3: A list of input arguments for the gca function

gca(Kmatrix, Xmatrix, sigma2a, sigma2e, MUScenario,

statistic, NumofMU, Uidx = NULL, scale = TRUE, diag = TRUE)

Example 1: Pairwise connectedness across units

The following example demonstrates the pairwise CD_IdAve across units with no additional fixed effect.

Box 4: Example of pairwise CD_IdAve across units

X_fixed <− **model.matrix**(∼ -1 + factor(cattle.pheno$Unit)) # incidence matrix of units

G <− **computeG**(cattle.W) # genomic relationship matrix

sigma2a <− 0.6 # additive genetic variance

sigma2e <− 0.4 # residual variance

CD_IdAve <− **gca**(Kmatrix = G, Xmatrix = X_fixed, sigma2a = sigma2a, sigma2e = sigma2e, MUScenario = **as.factor**(cattle.pheno$Unit), statistic = 'CD_IdAve', NumofMU = 'Pairwise')

Here, the 'X_fixed' is the incidence matrix of units with the intercept excluded. The 'G' is the first type of genomic relationship matrix in VanRaden [109]. The statistic 'CD_IdAve' calculates CD measures using individual average as a summary method. The option 'Pairwise' in the 'NumofMU' argument returns a square matrix containing pairwise connectedness measures across units.

Example 2: Overall connectedness across units

We present the calculation of overall CD_GrpAve measures across units by changing the argument of 'statistic' to 'CDGrpAve' in this example. The CD statistic is summarized at the unit level using PEV_{Mean} and PEC_{Mean}. Changing the argument 'NumofMU' to 'Overall' returns the average of all pairwise connectedness measures between units. The definitions of other arguments are identical as shown in Box 4.

Box 5: Example of overall CD_GrpAve across units

CD_GrpAve <− **gca**(Kmatrix = G, Xmatrix = X_fixed, sigma2a = sigma2a, sigma2e = sigma2e, MUScenario = **as.factor**(cattle.pheno$Unit), statistic = 'CD_GrpAve', NumofMU = 'Overall')

Example 3: Pairwise connectedness across units with fixed effects of units and sex

Box 6: Example of pairwise CDVED2 across units

X_fixed <− **model.matrix**(∼ -1 + factor(cattle.pheno$Unit)
+ **factor**(cattle.pheno$Sex)) # incidence matrix of units and sex
G <− **computeG**(cattle.W) # genomic relationship matrix
sigma2a <− 0.6 # additive genetic variance
sigma2e <− 0.4 # residual variance
CDVED2 <− **gca**(Kmatrix = G, Xmatrix = X_fixed, sigma2a = sigma2a, sigma2e = sigma2e, MUScenario = **as.factor**(cattle.pheno$Unit), statistic = 'CDVED2', NumofMU = 'Pairwise', Uidx = 8)

The above example shows the pairwise connectedness of CDVED2 while correcting for two fixed effects, namely units and sex. This code returns CD measures based on VE2.

5.4.5 Relationship between connectedness statistics

The relationships between connectedness statistics derived from PEV and VE have been probed by considerable studies in the past, and a highly correlated relationship has been reported [23, 102]. In this section, we specifically investigated the relationship between PEVD and VED, and between r and CR to compare with the results reported in the previous studies. The PEVD statistic is summarized using group average (PEVD_GrpAve) and the pairwise connectedness across units is reported in Figure 5A. Since the smaller PEVD refers to the greater connectedness, the most connected units were found between units 1 and 2 (PEVD_GrpAve = 0.0167), whereas units 4 and 6 (PEVD_GrpAve = 0.0571) showed the least connectedness. The statistics of variance of differences in unit effects (i.e., VED0, VED1, and VED2) were reported in Figure 5B, 5C and 5D, respectively. Analogous results were found in three VED statistics, where units 1, 2 has the greatest connectedness, and the least connectedness is found in units 4, 6. Furthermore, the corresponding correlation plots between PEVD and three VED statistics are showed in Figure 6. The statistic PEVD showed a strong relationship with three VED statistics. Increased correlations are identified between PEVD_GrpAve and VED0 (0.9996), VED1 (0.9999) and VED2 (1). A perfect positive correlation is identified between PEVD and VED2 when two fixed effects of unit effect and sex are corrected. These strong correlations between PEVD and VED are consistent with the results pointed out by Holmes et al [23]. Similar high correlations between group average r and connectedness rating (i.e., CR0, CR1, and CR2) are shown in Figure 7. These positive strong relationships are as expected according to the manner how fixed effects are accounted for by using corrected functions in the statistic of connectedness rating.

5.4.6 Relationship between connectedness metric and prediction accuracy

The following section illustrates the application of GCA package in estimating the degree of connectedness between testing and training sets under the whole-genome prediction framework. The testing and training sets are simulated by following Yu et al. [100]. All individuals are assigned to 8 clusters using the K-medoid algorithm [108], which aims to maximize the relationships between individuals within the cluster and minimize the relationships across the cluster simultaneously. Following this, the testing and training sets are formed in 5 ways. Eight clusters are randomly allocated to training and testing sets in Scenario 1, which is regarded as the least connected design. In Scenario 2 to 5, the degree of relatedness steadily increased by exchanging 140, 210, 280 and 350 randomly sampled individuals between training and testing sets in Scenario 1.

The connectedness between testing and training sets in 5 Scenarios is measured with statistics of PEVD_GrpAve and CD_GrpAve. Furthermore, the relationship between connectedness measure and prediction accuracy is investigated using a standard best linear unbiased prediction model under a Bayesian scheme following the procedure in Yu et al. [100]. The prediction accuracy is evaluated with two-fold cross-validation, where the predictive ability is assessed as the Pearson correlation between predicted breeding values and phenotypes in the testing set. The prediction accuracy changing with connectedness measure across 5 scenarios quantified with PEVD_GrpAve (first column) and CD_GrpAve (second column) is shown in Figure 8. The smaller PEVD_GrpAve and larger CD_GrpAve values suggest a stronger connectedness measure. As more individuals from the same cluster are shared between training and testing sets, increased connectedness measure and prediction accuracy are detected in the correlation between PEVD_GrpAve and PA. Whereas, CD_GrpAve

increased up to Scenario 3 and then decreased at Scenario 4, which acts to penalize the connectedness measure for the reduced genetic variability as suggested by previous studies [21, 100]. All these results are consistent with our previous study [100].

5.5 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. 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.

Figures

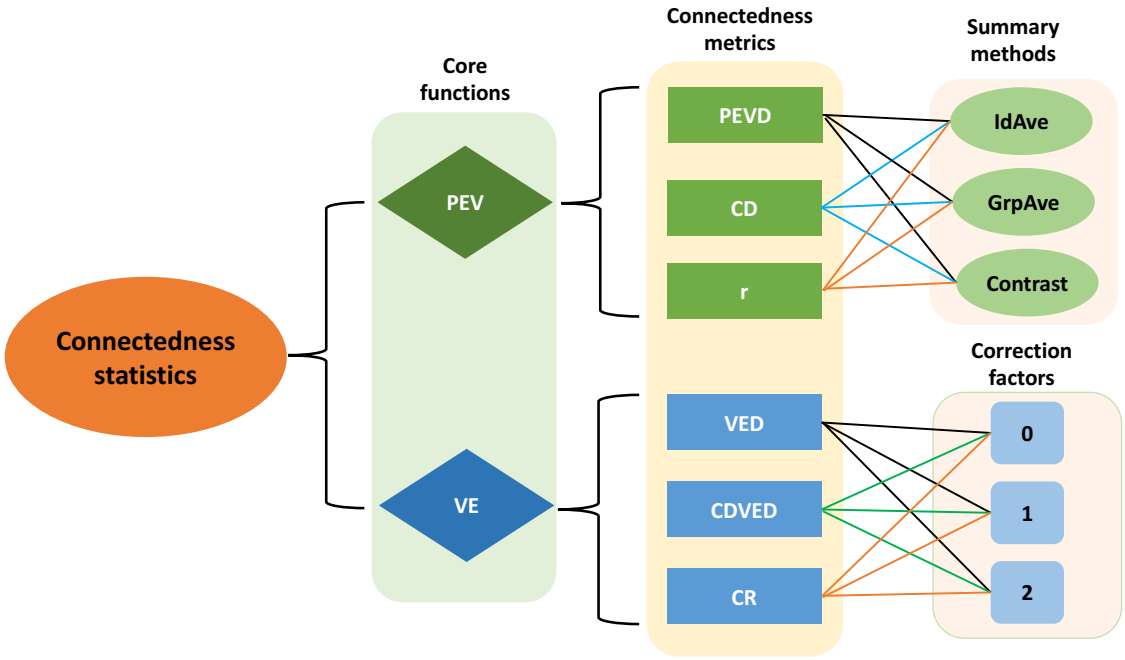

Figure 5.1: An overview of connectedness statistics implmented in the GCA R package. The statistics can be computed from either prediction error variance (PEV) or variance of unit effect estimates (VE). Connectedness metrics include prediction error variance of the difference (PEVD), coefficient of determination (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.

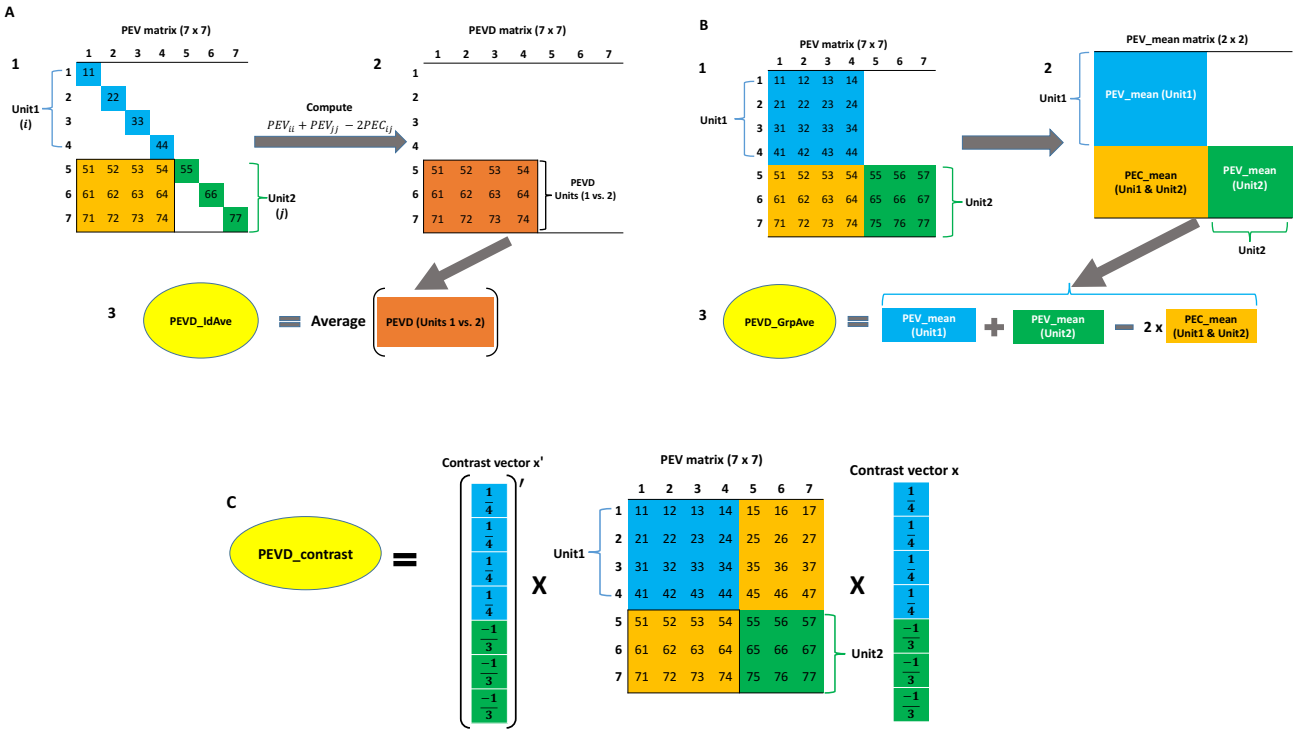

Figure 5.2: A flow diagram of three prediction error variance of the difference (PEVD) statistics. The individual average PEVD (PEVD_IdAve) is shown in A. A1: Prediction error variance (PEV) matrix including variances and covariances of seven individuals. Subscripts i and j refer to the ith and jth individuals in units 1 and 2, respectively. A2: Pairwise PEVD between individuals across two units. A3: Individual average PEVD is calculated by taking the average of all pairwise PEVD. The group average PEVD (PEVD_GrpAve) is shown in B. B1: Prediction error variance (PEV) matrix including variances and covariances of seven individuals. B2: Calculate the mean of prediction error variance /covariance within the unit (PEV_mean) and mean of prediction error covariance across the unit (PEC_mean). B3: Group average PEVD is calculated by applying the PEVD equation using PEV_mean and PEC_mean. 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.

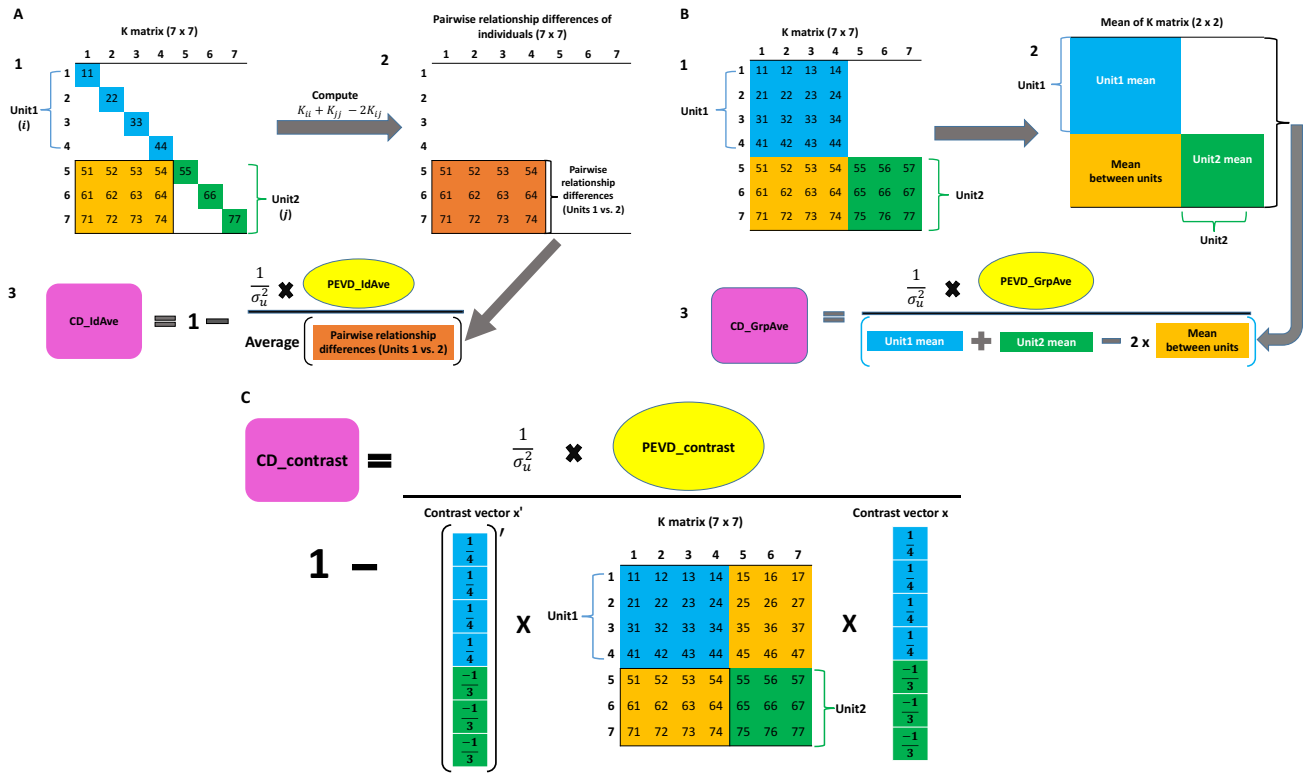

Figure 5.3: A flow diagram of three coefficient of determination (CD) statistics. The individual average CD (CD_IdAve) is shown in A. A1: A relationship matrix of seven individuals. A2: Calculate pairwise relationship differences of individuals between the units. Subscripts i and j refer to the ith and jth individuals in units 1 and 2, respectively. A3: Individual average CD is calculated by scaling indvidual average PEVD (PEVD_IdAve) with the average of pairwise relationship differences of individuals. The group average CD (CD_GrpAve) is shown in B. B1: A relationship matrix of seven individuals. B2: Calculate the mean relationships within and between units. B3: 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 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.

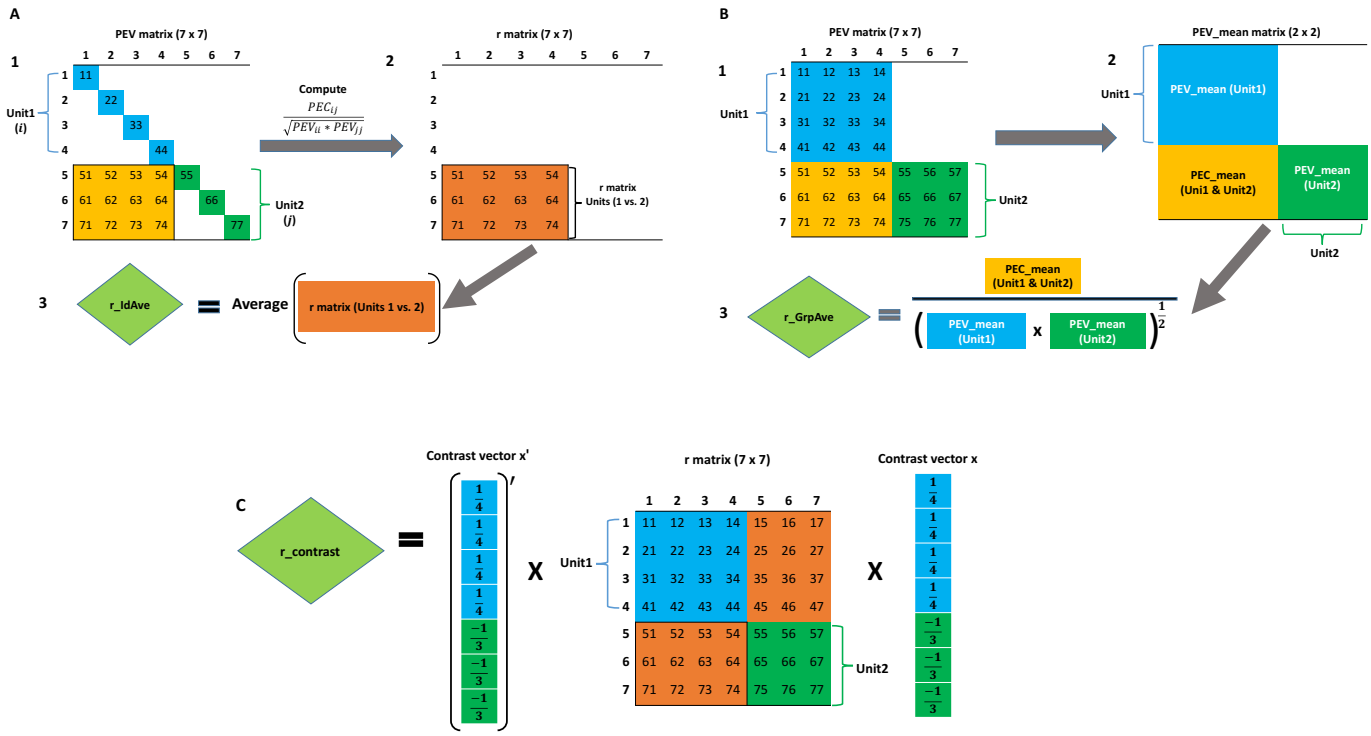

Figure 5.4: A flow diagram of three prediction error correlation (r) statistics. The calculation of individual average r (r_IdAve) involving seven individuals is displayed in A. A1: Prediction error variance (PEV) matrix of seven individuals. A2: Calculate pairwise correlation coefficients of individuals 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. A3: Individual average r is calculated as the average of pairwise prediction error correlation coefficients of individuals across units. The group average r (r_GrpAve) is shown in B. B1: Prediction error variance (PEV) matrix of seven individuals. B2: Calculate the mean of prediction error variance /covariance within the unit (PEV_mean) and mean of prediction error covariance across the unit (PEC_mean). B3: Group average r is a correlation calculated from PEV_mean and PEC_mean. 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.

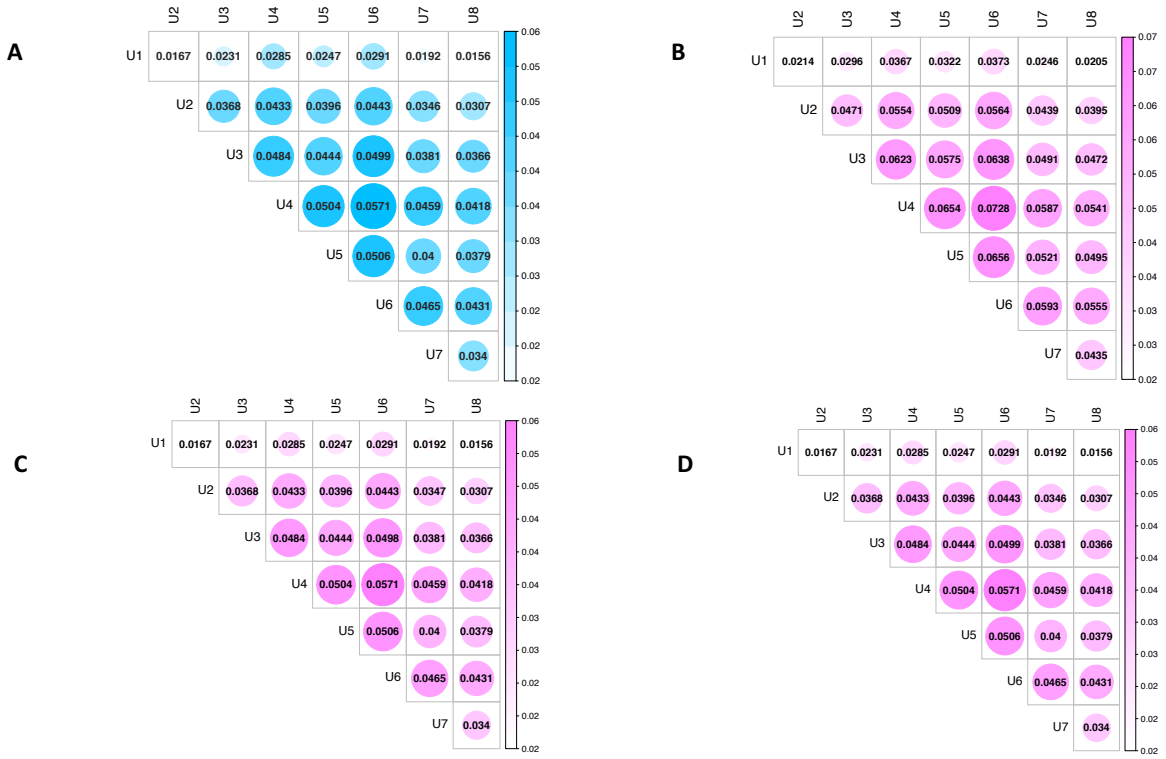

Figure 5.5: Pairwise connectedness across units. A: The group average PEVD (PEVD_GrpAve). B: Variance of differences in unit effects with no correction (VED0). C: Variance of differences in unit effects corrected unit effect (VED1). D: Variance of differences in unit effects with correction of all fixed effects (VED2)

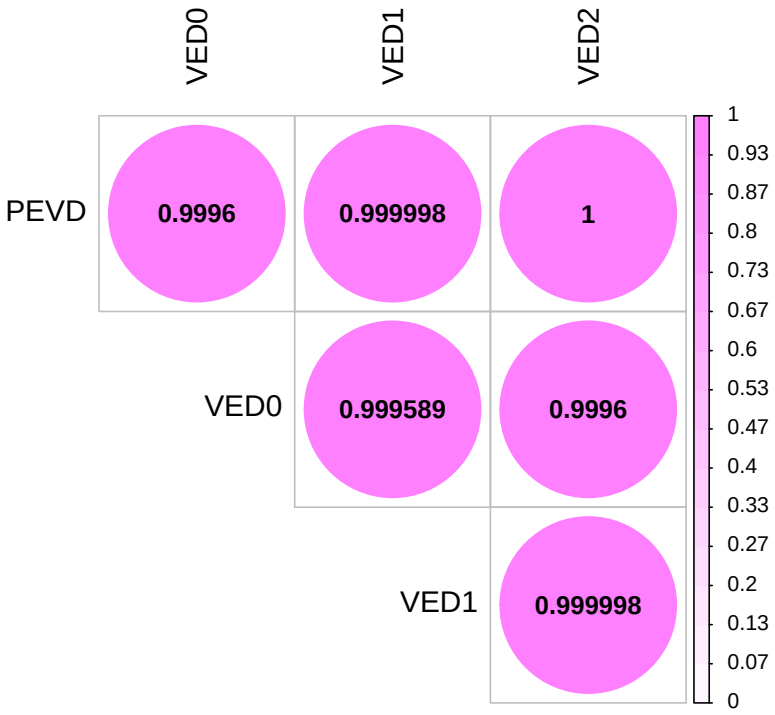

Figure 5.6: Heatmap to illustrate the correlation between the group average PEVD (PEVD_GrpAve) and VED0, VED1 and VED2.

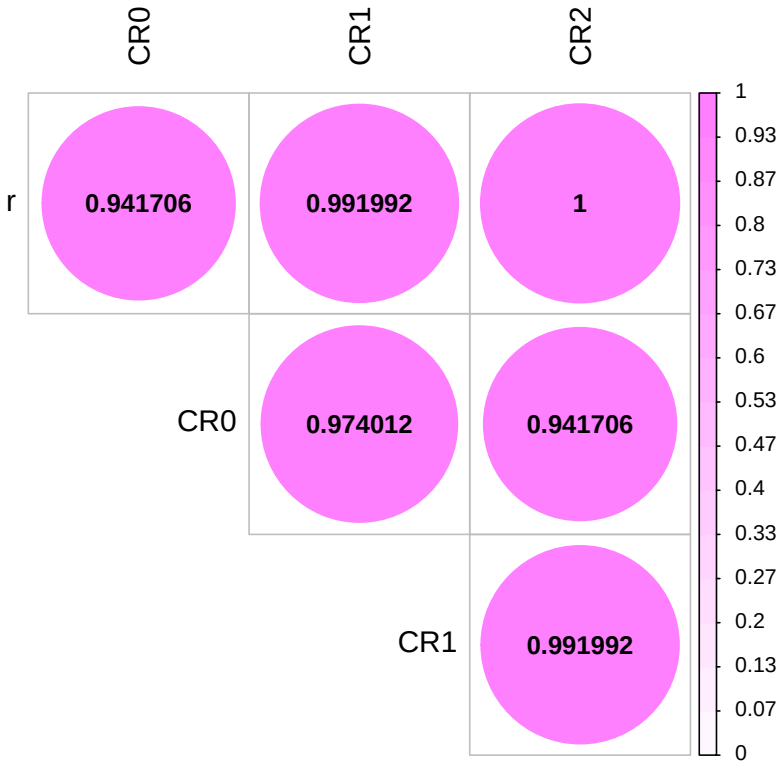

Figure 5.7: Heatmap to illustrate the correlation between the group average r (r_GrpAve) and CR0, CR1 and CR2.

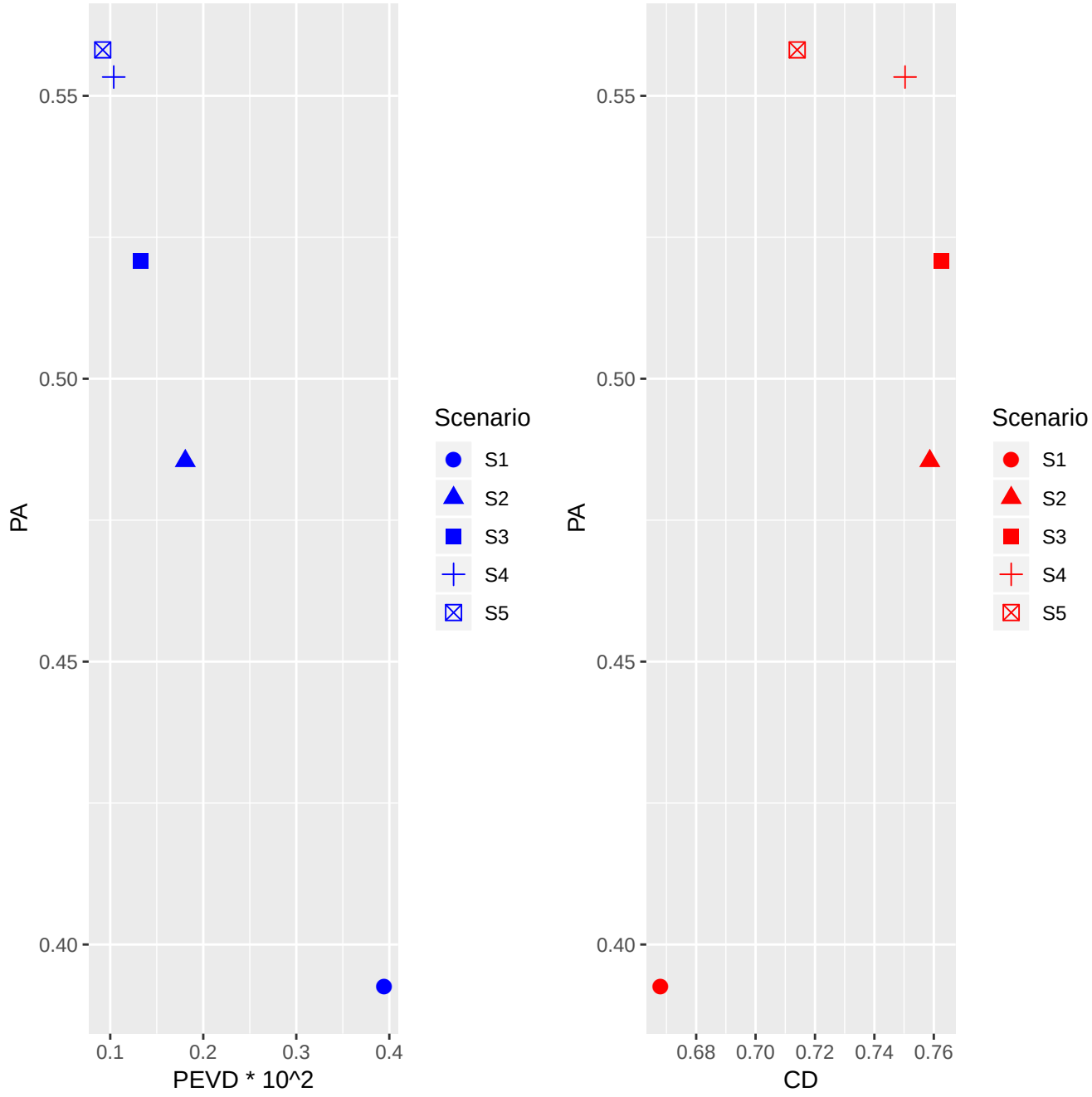

Figure 5.8: Relationship between connectedness and prediction accuracy across 5 scenarios. The PEVD and CD denote the group average PEVD and CD, respectively. The PA refers to the Pearson correlation between predicted breeding values and phenotypes in the testing set.

Chapter 6

Genomic Bayesian Confirmatory Factor Analysis and Bayesian Network To Characterize a Wide Spectrum of Rice Phenotypes

6.1 Abstract

With the advent of high-throughput phenotyping platforms, plant breeders have a means to assess many traits for large breeding populations. However, understanding the genetic interdependencies among high-dimensional traits in a statistically robust manner remains a major challenge. Since multiple phenotypes likely share mutual relationships, elucidating the interdependencies among economically important traits can better inform breeding decisions and accelerate the genetic improvement of plants. The objective of this study was to leverage confirmatory factor analysis and graphical modeling to elucidate the genetic interdependencies among a diverse agronomic traits in rice. We used a Bayesian network to depict conditional dependencies among phenotypes, which can not be obtained by standard multi-trait analysis. We utilized Bayesian confirmatory factor analysis which hypothesized that 48 observed phenotypes resulted from six latent variables including grain morphology, morphol-

ogy, flowering time, physiology, yield, and morphological salt response. This was followed by studying the genetics of each latent variable, which is also known as factor, using single nucleotide polymorphisms. Bayesian network structures involving the genomic component of six latent variables were established by fitting four algorithms (i.e., Hill Climbing, Tabu, Max-Min Hill Climbing, and General 2-Phase Restricted Maximization algorithms). Physiological components influenced the flowering time and grain morphology, and morphology and grain morphology influenced yield. In summary, we show the Bayesian network coupled with factor analysis can provide an effective approach to understand the interdependence patterns among phenotypes and to predict the potential influence of external interventions or selection related to target traits in the interrelated complex traits systems.

6.2 Introduction

A primary objective in plant breeding is to develop high yielding varieties with specific grain qualities, resilience to pests and abiotic stresses, and superior adaption to the target environment. As a result, plant breeders devote considerable resources to extensive phenotypic evaluation of germplasm and select on multiple traits. These traits are often correlated at a genetic level through common genetic effects (e.g., pleiotropy) or linkage disequilibrium between quantitative trait locus (QTL). Since multiple phenotypes may exhibit mutual relationships, knowledge of the interdependence among agronomically important traits can improve the efficacy of selection and rate of genetic improvement in systems with complex traits.

In a standard quantitative genetic analysis, multivariate phenotypes can be modeled through multi-trait models (MTM) of Henderson and Quaas [44] or some genomic counterparts [e.g., 46, 47] by leveraging genetic or environmental correlations among traits. In particular,

MTM has been useful in deriving genetic correlations and enhancing the prediction accuracy of breeding values for traits with low heritability or scarce records via joint modeling with one or more genetically correlated, highly heritable traits [45]. Conventional MTM strategies may provide important insight into the genetic relations between agronomically important traits, but they fail to explain how these traits are related. For instance, consider a case where we have three genetically correlated traits: y_1, y_2, and y_3. With MTM, we cannot address whether the relationship between y_1 and y_3 is due to direct effects, or if the relationship is driven by indirect effects mediated by y_2. Bayesian Networks (BN) offer an effective approach to elucidate the underlying network structure in multivariate data and infer network relationships between correlated variables. A BN is a probabilistic graphical model that represents conditional dependencies among a set of variables via a directed acyclic graph (DAG) [52]. In the DAG, the variables are represented by nodes, while their conditional dependencies between nodes are indicated with directed edges. In the context of plant breeding, BN can be used to elucidate the interdependencies among traits and inform selection decisions for simultaneously improving multiple traits. For instance in the latter case above ($y_1 \rightarrow y_2 \rightarrow y_3$), selection directly on y_2 will affect the quantity of y_3 without an effect on y_1.

With the advent of high-throughput phenotyping (HTP) platforms, plant breeders have been provided with a suite of tools for phenotypic evaluation of large populations [110]. These platforms leverage robotics, precise environmental control, and remote sensing techniques to provide accurate, repeatable and high resolution phenotypes for large breeding populations throughout the growing season [110, 111, 112]. These data can be used to redefine characteristics underlying superior agronomic performance by quantifying secondary traits associated with seedling vigor, plant architecture, photosynthesis, transpiration, disease resistance, and stress tolerance [5, 113, 114]. However given these new approaches, breeders are faced with

the new challenge of efficiently utilizing these large multidimesional data sets to improve selection efficiency. The primary challenges associated with multivariate analysis and BN approaches using HTP data is that robust parameter estimates can be untenable because the number of estimated parameters within the model increases with the increasing number of phenotypes. Moreover, even in cases where MTM or BN can be applied, interpreting of interrelationships among a large number of phenotypes can be difficult.

One approach to characterize high-dimensional phenotypes is by using factor analysis (FA). The central idea of FA approaches is to reduce the dimensions of multivariate data sets by constructing unobserved, latent factors, or modules, from correlated phenotypes [50]. The biological importance of these latent factors can be interpreted by inspecting the phenotypes that contribute to each factor. Thus, the advantage of FA for large, multivariate data sets is two fold. First, FA provides a means to reduce the dimensions of multivariate data sets thereby providing statistically sound parameter estimates, and easing visualization and interpretation. Secondly, the latent variables/factors themselves may be representative of underlying biological processes that cannot be observed or measured in the population. For instance, several studies have highlighted the effects of plant hormones such as GA on multiple morphological attributes [115, 116, 117, 118, 119, 120]. Thus, a latent factor constructed from these morphological traits may provide information on the biosynthesis or sensitivity of these hormones for individuals within the population. If a certain amount of knowledge regarding the biological role of the variables is already known, a varaint of FA, confirmatory factor analysis (CFA), can be used to estimate latent variables based on predetermined biological classes of observed traits [121]. These latent variables underlie observed phenotypes and can be evaluated for how well the data support the hypothesis. For instance, Peñagaricano et al. [51] performed CFA in swine to derive five latent variables from 19 phenotypic traits and inferred BN structures among those latent variables, thereby

demonstrating the potential of this approach.

This study aimed to leverage CFA and graphical modeling to elucidate the genetic interdependencies among traits typically recorded in breeding programs (e.g., yield, plant morphology, phenology, and stress resilience). First, we constructed latent variables, using prior biological knowledge obtained from the literature. Then we connected the observed high-dimensional phenotypes with these to establish latent variables via Bayesian confirmatory factor analysis (BCFA) to reduce the dimensions of the dataset. Further, factor scores computed from BCFA were considered new phenotypes for a Bayesian multivariate analysis to separate breeding values from noise. This was followed by adjustment of breeding values via Cholesky decomposition to eliminate the dependencies introduced by genomic relationships. Finally, the adjusted breeding values were considered inputs to assess the network structure between latent variables by conducting a Gaussian BN analysis. This study is the first, to our knowledge, in rice to characterize various phenotypes with graphical modeling such as BCFA and BN.

6.3 Materials and Methods

6.3.1 Phenotypic and genotypic data

The rice dataset comprised $n = 374$ accessions sampled from six subpopulations: temperate japonica (92), tropical japonica (85), indica (77), aus (52), aromatic (12), and admixture of japonica and indica (56) [122]. The improvement status of each accession was obtained from the USDA-ARS Germplasm Resources Information Network. We used $t = 48$ phenotypes and data regarding 44,000 single-nucleotide polymorphisms (SNP). After removing SNP markers with minor allele frequency less than 0.05, 374 accessions and 33,584 markers were

used for further analysis. Of those, 27 phenotypes were reported in Zhao et al. [122] and McCouch et al. [123]. These phenotypes can be classified into four categories: flowering time (flowering time at three locations, photoperiod sensitivity), grain morphology (seed length, seed width, seed surface area, seed length to width ratio, seed volume), plant morphology (culm habit/angle, flag leaf length and width, plant height at maturity), and yield traits (panicle fertility, seed number per panicle, number of primary branches on the main panicle, panicle length, and the number of panicles on each plant). Zhao et al. [122] evaluated flowering time-related traits using data from three locations, while the remaining traits were evaluated at one location (Arkansas). The remaining phenotypes were assessed from the salinity stress experiments conducted in Campbell et al. [124]. These traits were classified into three categories: morphological salt response, ionic components of salt stress, and plant morphology. The class morphological salt response represents how plant growth is affected by salinity stress and is composed of the ratio of shoot biomass of salt stressed plants to control, the ratio of root biomass of salt stressed plants to control, the ratio of the number of tillers for salt stressed plants to control, and two metrics that represent the ratio of shoot height of salt stressed plants to control. Ionic components of salt stress is composed of traits that quantify ions important for salinity tolerance (Na^+ and K^+) in both root and shoot tissues. Morphology traits are those that describe the growth of the plant in both control and saline conditions (e.g. shoot biomass, root biomass, shoot height, and tiller number). The data used from Campbell et al. [124] were derived from three to six independent greenhouse experiments performed between July and October 2013. Information for all experiments were combined and best linear unbiased estimators were calculated for each line as described in Campbell et al. [124]. The detailed descriptions of the phenotypes are summarized in Appendix C.

6.3.2 Bayesian confirmatory factor analysis

A CFA under the Bayesian framework was performed to model 48 phenotypes. The number of factors and the pattern of phenotype-factor relationships need to be specified in BCFA prior to model fitting. We constructed six latent variables ($q = 6$) from previous reports [122, 124, 125]. The six latent variables derived from our analysis represent the grain morphology, morphology, flowering time, ionic components of salt stress, yield, and morphological salt response (Appendix C). Each latent variable captures common signals spanning genetic and environmental effects across all its phenotypes. The latent variables, which determine the observed phenotypes can be modeled as

$$\mathbf{T} = \mathbf{\Lambda F} + \mathbf{s},$$

where $\mathbf{T}$ is the $t \times n$ matrix of observed phenotypes, is the $t \times q$ factor loading matrix, $\mathbf{F}$ is the $q \times n$ latent variables matrix, and $\mathbf{s}$ is the $t \times n$ matrix of specific effects. Here, maps latent variables to the observed variables and can be interpreted as the extent of contribution each latent variable to phenotype. This can be derived by solving the following variance-covariance model.

$$var(\mathbf{T}) = \mathbf{\Lambda \Phi \Lambda}' + \mathbf{\Psi},$$

where $\mathbf{\Phi}$ is the variance of latent variables, and $\mathbf{\Psi}$ is the variance of specific effects [126]. Six latent variables were assumed to account for the covariance in the observed phenotypes. Moreover, latent variables were assumed to be correlated with each other. Prior distributions were assigned to all unknown parameters. The non-zero coefficients within factor loading matrix $\mathbf{\Lambda}$ were assumed to follow a Gaussian distribution with mean of 0 and variance of 0.01. The variance-covariance matrix $\mathbf{\Phi}$ was assigned an inverse Wishart distribution with

a 6×6 identity scale matrix $\mathbf{I}_{66}$ and a degree of freedom 7, $\mathbf{\Phi} \sim \mathcal{W}^{-1}(\mathbf{I}_{66}, 7)$ and an inverse Gamma distribution with scale parameter 1 and shape parameter 0.5 was assigned to $\mathbf{\Psi} \sim \Gamma^{-1}(1, 0.5)$.

We employed the blavaan R package [127] jointly with JAGS [128] to fit the above BCFA. The blavaan runs the runjags R package [129] to summarize the Markov chain Monte Carlo (MCMC) and samples unknown parameters from the posterior distributions. Three MCMC chains, each of 5,000 samples with 2,000 burn-in, were used to infer the unknown model parameters. The convergence of the parameters was investigated with trace plots and potential scale reduction factor (PSRF) less than 1.2 [130]. The PSRF computes the difference between estimated variances among multiple Markov chains and estimated variances within the chain. A large difference indicates non-convergence and may require additional Gibbs sampling.

Subsequently, the posterior means of factor scores ($\mathbf{F}$), which reflect the contribution of latent variables to each accession were estimated. Within each draw of Gibbs sampling, $\mathbf{F}$ was sampled from the conditional distribution of $p(\mathbf{F}|\boldsymbol{\theta}, \mathbf{T})$, where $\boldsymbol{\theta}$ refers to the unknown parameters in $\mathbf{\Lambda}$, $\mathbf{\Phi}$, and $\mathbf{\Psi}$. This conditional distribution was derived with data augmentation [131] assuming $\mathbf{F}$ as missing data [132].

6.3.3 Multivariate genomic best linear unbiased prediction

We fitted a Bayesian multivariate genomic best linear unbiased prediction to separate breeding values from population structure and noise in the six factor scores computed previously.

$$\mathbf{F} = \boldsymbol{\mu} + \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\mu}$ is the vector of intercept, $\mathbf{X}$ is the incidence matrix of covariates, $\mathbf{b}$ is the vector of covariate effects, $\mathbf{Z}$ is the incidence matrix relating accessions with additive genetic effects, $\mathbf{u}$ is the vector of additive genetic effects, and $\boldsymbol{\epsilon}$ is the vector of residuals. The incident matrix $\mathbf{X}$ included subpopulation information (temperate japonica, tropical japonica, indica, aus, aromatic, and admixture), as the rice diversity panel used herein shows a clear substructure [122].

A flat prior was assigned to $\boldsymbol{\mu}$ and $\mathbf{b}$, and the joint distribution of $\mathbf{u}$ and $\boldsymbol{\epsilon}$ follows multivariate normal

$$\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N \left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{u}} \otimes \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \mathbf{I} \end{pmatrix} \right],$$

where $\mathbf{G}$ represents the second genomic relationship matrix of VanRaden [69], $\mathbf{I}$ is the identity matrix, $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ refer to 6×6 dimensional genetic and residual variance-covariance matrices, respectively. An inverse Wishart distribution with a 6×6 identity scale matrix of $\mathbf{I}_{66}$ and a degree of freedom 6 was assigned as prior for $\boldsymbol{\Sigma}_{\boldsymbol{u}}, \boldsymbol{\Sigma}_{\boldsymbol{e}} \sim \mathcal{W}^{-1}(\mathbf{I}_{66}, 6)$. These parameters were selected so that relatively uninformative priors were used. The Bayesian multivariate genomic best linear unbiased prediction model was implemented using the MTM R package (https://github.com/QuantGen/MTM). Posterior mean estimates of genomic correlation between latent variables and predicted breeding values ($\widehat{\mathbf{u}}$) were then obtained. The convergence of the estimated parameters was verified by trace plots.

6.3.4 Sample independence in the Bayesian network

Theoretically, BN learning algorithms assume sample independence. In the multivariate genomic best linear unbiased prediction, the residuals between phenotypes were assumed

independent through $\mathbf{I_{374x374}}$. However, phenotypic dependencies were introduced by the $\mathbf{G}$ matrix for the additive genetic effects, thereby potentially serving as a confounder. Thus, a transformation of $\widehat{\mathbf{u}}$ was carried out to derive an adjusted $\widehat{\mathbf{u}}^*$ by eliminating the dependencies in $\mathbf{G}$. For a single trait model, the adjusted $\widehat{\mathbf{u}}^*$ can be computed by premultiplying $\widehat{\mathbf{u}}$ by $\mathbf{L}^{-1}$, where $\mathbf{L}$ is a lower triangular matrix derived from the Choleskey decompostion of $\mathbf{G}$ matrix ($\mathbf{G} = \mathbf{L}\mathbf{L}^{'}$). Since $\mathbf{u} \sim \mathcal{N}(0, \mathbf{G}\sigma_u^2)$, the distribution of $\widehat{\mathbf{u}}^*$ follows $\mathcal{N}(0, \mathbf{I}\sigma_u^2)$ [133, 134]

$$\begin{aligned} Var(\mathbf{u}^*) &= Var(\mathbf{L^{-1}u}) \\ &= \mathbf{L^{-1}}Var(\mathbf{u})(\mathbf{L^{-1}})^{'} \\ &= \mathbf{L^{-1}G}(\mathbf{L^{-1}})^{'}\sigma_u^2 \\ &= \mathbf{L^{-1}L}\mathbf{L}^{'}(\mathbf{L}')^{-1}\sigma_u^2 \\ &= \mathbf{I}\sigma_u^2. \end{aligned}$$

This transformation can be extended to a multi-traits model by defining $\mathbf{u}^* = \mathbf{M^{-1}u}$, where $\mathbf{M^{-1}} = \mathbf{I_{qq}} \otimes \mathbf{L^{-1}}$ [55]. Under the multivariate framework, $\mathbf{u}$ follows $\mathcal{N}(0, \boldsymbol{\Sigma_u} \otimes \mathbf{G})$ and the variance of $\mathbf{u}^*$ is

$$\begin{aligned} Var(\mathbf{u}^*) &= Var(\mathbf{M^{-1}u}) \\ &= (\mathbf{I_{qq}} \otimes \mathbf{L^{-1}})(\boldsymbol{\Sigma_u} \otimes \mathbf{G})(\mathbf{I_{qq}} \otimes \mathbf{L^{-1}})^{'} \\ &= (\mathbf{I_{qq}} \otimes \mathbf{L^{-1}})(\boldsymbol{\Sigma_u} \otimes \mathbf{L}\mathbf{L}^{'})(\mathbf{I_{qq}} \otimes \mathbf{L^{-1}})^{'} \\ &= \boldsymbol{\Sigma_u} \otimes \mathbf{I_{nn}}, \end{aligned}$$

where $\mathbf{L^{-1}L}\mathbf{L}^{'}(\mathbf{L^{-1}})^{'} = \mathbf{I_{nn}}$. This adjusted $\widehat{\mathbf{u}}^*$ was used to learn BN structures between predicted breeding values.

6.3.5 Bayesian network

A BN depicts the joint probabilistic distribution of random variables through their conditional independencies [135]

$$\mathcal{BN} = (\mathcal{G}, X_V),$$

where $\mathcal{G}$ represents a DAG = (V, E) with nodes (V) connected by one or more edges (E) conveying the probabilistic relationships and the random vector $X_V = (X_1, ..., X_K)$ is K random variables. The joint probability distribution can be factorized as

$$P(X_V) = P(X_1, ..., X_K) = \prod_{v=1}^{K} P(X_v|Pa(X_v)),$$

where $Pa(X_v)$ denotes a set of parent nodes of child node X_v. The DAG and joint probability distribution are governed by the Markov condition, which states that every random variable is independent of its non-descendants conditioned on its parents. A BN is known as a Gaussian BN, when all variables or phenotypes are defined as marginal or conditional Gaussian distribution as in the present study.

The adjusted breeding values $\widehat{\mathbf{u}}^*$ were used to infer a genomic network structure among the aforementioned six latent variables. There are three types of structure-learning algorithms for BN: constraint-based algorithms, score-based algorithms, and a hybrid of these two [135]. The constraint-based algorithms can be originally traced to the inductive causation algorithm [136], which uses conditional independence tests for network inference. Briefly, the first step is to identify a d-separation set for each pair of nodes and confer an undirected edge between the two if they are not d-separated. The second step is to identify a v-structure for each pair of non-adjacent nodes, where a common neighbor is the outcome of two non-adjacent nodes.

In the last step, compelled edges were identified and oriented, where neither cyclic graph nor new v-structures are permitted. The score-based algorithms are based on heuristic approaches, which first assign a goodness-of-fit score for an initial graph structure and then maximize this score by updating the structure (i.e., add, delete, or reverse the edges of initial graph). The hybrid algorithm includes two steps, restrict and maximize, which harness both constraint-based and score-based algorithms to construct a reliable network. In this study, the two score-based (Hill Climbing and Tabu) and two hybrid algorithms (Max-Min Hill Climbing and General 2-Phase Restricted Maximization) were used to perform structure learning. A flow diagram to illustrate the concept of constraint-based Bayesian netwrok structure learning algorithm is shown in Figure 6.1.

We quantified the strength of edges and uncertainty regarding the direction of networks, using 500 bootstrapping replicates with a size equal to the number of accessions and performed structure learning for each replicate in accordance with Scutari and Denis [135]. Non-parametric bootstrap resampling aimed at reducing the impact of the local optimal structures by computing the probability of the arcs and directions. Subsequently, 500 learned structures were averaged with a strength threshold of 85% or higher to produce a more robust network structure. This process, known as model averaging, returns the final network with arcs present in at least 85% among all 500 networks. Candidate networks were compared on the basis of the Bayesian information criterion (BIC) and Bayesian Gaussian equivalent score (BGe). The BIC accounts for the goodness-of-fit and model complexity, and BGe aims at maximizing the posterior probability of networks per the data. All BN were learned via the bnlearn R package [137]. In bnlearn, the BIC score is rescaled by -2, which indicates that the larger BIC refers to a preferred model.

6.4 Results

To elucidate the genetic interdependencies among traits typically recorded in breeding programs, we utilized a collection of 48 publicly available phenotypes recorded on a panel of diverse rice accessions [122, 124]. The phenotypic data was derived from two independent studies. The first set of phenotypes was recorded from materials grown in two field environments in Arkansas and Faridpur Bangladesh, and in a greenhouse in Aberdeen, UK [122]. The 34 phenotypes were recorded at maturity and were largely associated with yield (panicle characteristics flowering time, plant morphology (e.g., height and growth habits), and seed morphological traits. The second study consisted of 14 phenotypes were recorded in a greenhouse environment on plants in the active tillering stage (e.g., 30 day-old plants) under control and saline (14 days of 9.5 dS m−2 NaCl stress). The phenotypes from this study can be classified into three categories: morphological traits (e.g., shoot and root biomass, and plant height), morphological responses to salinity (e.g., the ratio of morphological traits in saline conditions to control), and the ionic components of salinity stress (e.g., Na^+, K^+, and Na^+:K^+ in both root and shoot tissues) [124]. The complete data set provides an in-depth characterization of phenotypic performance at vegetative and reproductive stages in rice using several classes of traits.

6.4.1 Latent variable modeling

The BCFA model grouped the observed phenotypes into the underlying latent variables on the basis of prior biological knowledge, assuming these latent variables determine the observed phenotypes. This allowed us to study the genetics of each latent variable. A measurement model derived from BCFA evaluating the six latent variables is shown in Figure 6.2. Forty-eight observed phenotypes were hypothesized to result from the six latent variables: 7

for flowering time, 14 for morphology, 5 for yield, 11 for grain morphology, 6 for physiology, and 5 for salt response. The convergence of the parameters was confirmed graphically with the trace plots and a PSRF value less than 1.2 [127, 130].

The six latent factors showed strong contributions to the 48 observed phenotypes, with standardized regression coefficients ranging from -0.549 to 0.990 for flowering time, -0.349 to 0.925 for morphology, -0.085 to 0.790 for yield, -0.476 to 0.990 for grain morphology, -0.265 to 0.983 for ionic components of salt stress, and -0.022 to 0.939 for salt response. The latent factor flowering time showed a strong positive contribution to flowering time in Arkansas (Fla) and Flowering time in Arkansas in 2007 (Fla7) (0.990 and 0.926, respectively; Table 6.1), indicating that larger values for the latent factor can be interpreted as a greater number of days from sowing to emergence of the inflorescence. The latent factor morphology showed the largest positive contributions to traits describing height during the vegetative stage (e.g., height to newest ligule in salt (Hls), 0.920; height to newest ligule in control (Hlc), 0.899; height to the tip of first fully expanded leaf in salt (Hfs), 0.907; and height to tip of first fully expanded leaf in control (Hfc), 0.925;) suggesting that this latent factor is an overall representation of plant size. Yield showed large positive contributions to the observed phenotypes primary panicle branch number (Ppn) and seed number per panicle (Snpp) (0.790 and 0.780, respectively), suggesting that larger values for yield indicate a higher degree of branching and seed number. Observed phenotypes describing seed size (e.g., seed volume (Sv) and brown rice volume (Bvl) (0.990 and 0.986, respectively)) were most strongly associated with grain morphology. The latent factor ionic components of salt stress showed strong positive contributions to two observed phenotypes that quantify the ionic components of salt stress (shoot Na^+:K^+ (Ks) and shoot Na^+ (Nas) (0.983 and 0.975, respectively), indicating that higher values for the latent factor result in greater shoot Na^+ and Na^+:K^+. Finally, the latent factor describing morphological salt response showed strong

positive contributions to the observed phenotype describing the effect of salt treatment on plant height (ratio of height to tip of newest fully expanded leaf in salt to that of control plants (Hfr) (0.939)), thus larger values for the latent factor may indicate a more tolerant growth response to salinity.

6.4.2 Genomic correlation among latent variables

To understand the genetic relationships between latent variables, genomic correlation analysis was performed. Genomic correlation is due to pleiotropy or linkage disequilibrium between QTL. The genomic correlations among latent variables are shown in Figure 6.3. Negative correlations were observed between morphological salt response (Msr) and all other five latent variables. In particular, flowering time (-0.5), yield (-0.54), and grain morphology (-0.74) were negatively correlated with morphological salt response. These results suggest that accessions that harbor alleles for more tolerant morphological salt responses may also have alleles associated with longer flowering times, smaller seeds, and low yield. Similarly, a negative correlation was observed between morphology and yield (-0.56) and between morphology and grain morphology (-0.31). Thus, accessions with alleles associated with large plant size may also have alleles that result in low yield, small grain volume, and lower shoot Na^+ and Na^+:K^+. In contrast, a positive correlation was observed between grain morphology and yield (0.49) and between grain morphology and ionic components of salt stress (0.4). Thus, selection for large grain may result in improved yield, and higher shoot Na^+ and Na^+:K^+.

6.4.3 Bayesian network

To infer the possible network structure between latent variables, BN was performed. Prior to BN, the normality of latent variables was assessed using histogram plots combined with

density curves as shown in Supplementary Figure 6.4. Overall, all the six latent variables approximately followed a Gaussian distribution.

The Bayesian networks learned with the score-based and hybrid algorithms are shown in Figure 6.5. The structures of BN were refined by model averaging with 500 networks from bootstrap resampling to reduce the impact of local optimal structures. The labels of the arcs measure the uncertainty of the arcs, corresponding to strength and direction (in parenthesis). The former measures the frequency of the arc presented among all 500 networks from the bootstrapping replicates and the latter is the frequency of the direction shown conditional on the presence of the arc. We observed minor differences in the structures presented within and across the two types of algorithms used. In general, small differences were observed within algorithm types compared to those across algorithms. The two score-based algorithms produced a greater number of edges than two hybrid algorithms. The Hill Climbing algorithm produced seven directed connections among the six latent variables. Three connections were indicated towards flowering time from morphological salt response, ionic components of salt stress, and morphology, and two edges to yield from morphology and from grain morphology. Other two edges were observed from ionic components of salt stress to grain morphology and from grain morphology to morphological salt response. A similar structure was generated by the Tabu algorithm, except that the connection between salt response and grain morphology presented an opposite direction. The Max-Min Hill Climbing hybrid algorithm yielded six directed edges from morphological salt response to grain morphology, from ionic components of salt stress to grain morphology, from ionic components of salt stress to flowering time, from flowering time to morphology, from morphology to yield, and from grain morphology to yield. An analogous structure with the only difference observed in the directed edge from morphology to flowering time was inferred with the General 2-Phase Restricted Maximization algorithm. Across all four algorithms, there were four common directed edges: from ionic

components of salt stress to flowering time and to grain morphology, and from morphology and grain morphology to yield. The most favorable network was considered the one from the Tabu algorithm, which returned the largest network score in terms of BIC (1086.61) and BGe (1080.88). Collectively, these results suggest that there may be a direct genetic influence of morphology and grain morphology on yield, and physiological components of salt tolerance on grain morphology and flowering time.

6.5 Discussion

This study is based on the premise that most phenotypes interact to greater or lesser degrees with each other through underlying physiological and molecular pathways. While these physiological pathways are important for the development of agronomically important characteristics, they are often unknown or difficult to assess in large populations. The approach utilized here leverages phenotypes that can be readily assessed in large populations to quantify these underlying unobserved phenotypes, and elucidates the relationships between these variables.

Understanding the behaviors among phenotypes in the complex traits is critical for genetic improvement of agricultural species [138]. Graphical modeling offers an avenue to decipher bi-directional associations or probabilistic dependencies among variables of interest in plant and animal breeding. For instance, BN and L1-regularized undirected network can be used to model interrelationships of linkage disequilibrium (LD) [53, 139] or phenotypic, genetic, and environmental interactions [54] in a systematic manner. Importantly, MTM elucidates both direct and indirect relationships among phenotypes. Inaccurate interpretation of these relationships may substantially bias selection decisions [140, 141]. Thus, we applied BCFA to reduce the dimension of the responses by hypothesizing 48 manifest phenotypes originated

from the underlying six constructed latent variables as shown in Figure 6.2 assuming that these latent traits are most important, followed by application of BN to infer the structures among the six biologically relevant latent variables (Figure 6.5). Note that there are two differences between the approach employed here and a path analysis. A path analysis 1) uses observed variables rather than latent variables and 2) assumes a network structure is known priori. Thus, one advantage of our approach is that it can model a network structure at the level of latent variables and infer a network structure directly from data when prior information is not available from the literature or previous experiments. The BN represents the conditional dependencies between variables. Care must be taken in interpreting these relationships as a causal effect. Although a good BN is expected to describe the underlying causal structure per the data, when the structure is learned solely on the basis of the observed data, it may return multiple equivalent networks that describe the data well. In practice, searching such a causal structure with observed data needs three additional assumptions [135]: 1) each variable is independent of its non-effects (i.e., direct and indirect) conditioned on its direct causes, 2) the probability distribution of variables is supported by a DAG, where the d-separation in DAG provides all dependencies in the probability distribution, and 3) no additional variables influence the variables within the network. Although it may be difficult to meet these assumptions in the observed data, a BN is equipped with suggesting potential causal relationships among latent variables, which can assist in exploring data, making breeding decisions, and improving management strategies in breeding programs [142].

6.5.1 Biological meaning of latent variables and their relationships

We performed BCFA to summarize the original 48 phenotypes with the six latent variables. The number of latent variables and which latent variables load onto phenotypes were determined from the literature. The latent variable morphological salt response (Msr) con-

tributed strongly to salt indices for shoot biomass, root biomass, and two indices for plant height (Table 6.1). Thus, morphological salt response can be interpreted as the morphological responses to salinity stress, with higher values indicating a more tolerant growth response. The latent variable yield is a representation of overall grain productivity, and contributed strongly to the observed phenotypes primary panicle branch number, seed number per panicle, and panicle length. The positive loading scores on these observable phenotypes indicates that more highly branched, productive panicles will have higher values for yield (Table 6.1). Seed width, seed volume, and seed surface area contributed significantly to the latent variable grain morphology (Grm) (Table 6.1). Therefore, these results indicate that the grain morphology is a summary of the overall shape of the grain, where high values represent large, round grains, while low values represent small, slender grains. Considering the grain characteristics of rice subpopulations, temperate japonica accessions are expected to have high values for grain morphology, while indica accessions have lower values for grain morphology. Latent variable morphology (Mrp) is a representation of plant biomass during the vegetative stage (28-day-old plants) (Table 6.1). Shoot biomass, root biomass, and two metrics for plant height contributed largely to morphology, suggesting that accessions with high values for morphology are tall plants with a large biomass.

Genomic correlation analysis among the six latent variables showed meaningful correlations among several pairs. These genetic correlations can either be caused by linkage or pleiotropy. The former is likely to prevail in species with high LD, which is the case in rice where LD ranges from 100 to 200kb [143]. A negative relationship was observed between morphological salt response and three other latent variables (Figure 6.3). For instance, a negative correlation between morphological salt response and yield indicates that accessions of samples harboring alleles for superior morphological salt responses (e.g., those that are more tolerant) tend to also harbor alleles for poor yield (Figure 6.3). The rice diversity panel we used

is a representative sample of the total genetic diversity within cultivated rice and contains many unimproved traditional varieties (∼12% of lines in the study are landraces and ∼33% classified as cultivars; Supplementary File S2) and modern breeding lines [144]. While traditional varieties exhibit superior adaptation to abiotic stresses, they often have very poor agronomic characteristics including low yield, late flowering, and high photoperiod sensitivity [145, 146]. Moreover, the indica and japonica subspecies have contrasting salt responses and very different grain morphology. Japonica accessions tend to have short, round seeds and are more sensitive to salt stress, while indica accessions have long, slender grains and often are more salt tolerant [122, 124]. The negative relationship observed between morphological salt response and grain morphology suggests that lines that harbor alleles for high grain morphology (e.g., large, round grains) tend to also harbor alleles for a tolerant growth response to salt stress. However, no studies have yet reported an association between alleles for grain morphology and morphological salt response. Therefore, it remains to be addressed whether this relationship is due to LD or pleitropy.

Genetic correlations observed between other latent variables may suggest a pleiotropic effect among loci. For instance, a negative relationship was observed between morphological salt response and ionic components of salt stress, indicating that accessions harboring alleles associated with superior morphological salt response also tend to harbor alleles for reduced ion content under salt stress (Figure 6.3). The relationship between salt tolerance, measured in terms of growth or yield, and Na^+ and Na^+:K^+ has been a documented for decades (reviewed by [147]). Moreover, natural variation for Na^+ transporters has been utilized to improve growth and yield under saline conditions in rice and other cereals [124, 148, 149, 150, 151]. Therefore, the negative genetic relationships observed between morphological salt response and ion content may be due to the pleiotropic effects of some loci.

The genomic relationships among latent variables including morphology, yield, and grain

morphology may have resulted from the selection of alleles associated with good agronomic characteristics. A positive relationship was observed between yield and grain morphology, suggesting that alleles that positively contribute to productive panicles also may contribute to large, round grains. Furthermore, the negative genomic correlation observed between morphology and yield indicates that alleles negatively influencing total plant biomass also have a positive contribution to traits for productive panicles. This genomic relationship may reflect the genetics of harvest index, which is defined as the ratio of grain yield to total biomass. Over the past 50 years, rice breeders have selected high harvest index, resulting in plants with short compact morphology and many highly productive panicles [152, 153].

Although BCFA may yield biologically meaningful results, a potential limitation of BCFA is that we assumed each phenotype does not measure more than one latent variable. This assumption may not always strictly concur with the observational data. Therefore, further studies are required to allow each phenotype to potentially load onto multiple factors in the BCFA framework. An alternative approach is to derive the number of latent variables and determine which latent variables load onto phenotypes directly from observed data, using exploratory FA. This approach was not pursued here because accurate estimation of unknown parameters in the exploratory FA requires a large sample size, which was not the case herein [126].

6.5.2 Bayesian network of latent variables

The BN is a probabilistic DAG, which represents the conditional dependencies among phenotypes. The genomic correlation among latent variables described in Figure 6.3 does not inform the flow of genetic signals nor distinguish direct and indirect associations, whereas BN displays directions between latent variables and separate direct and indirect associations.

Therefore, the BN describes the possibility that other phenotypes will change if one phenotype is intervened (i.e., selection). However, caution is required to interpret this network as a causal effect, as the causal BN requires more assumptions, which are usually difficult to meet in observational data [154].

Four common edges or consensus subnetworks across the four BN may be the most reliable substructure of latent variables and may describe the dependence between agronomic traits (Figure 6.5). For example, edges from grain morphology to yield and morphology to yield can be interpreted as final grain productivity is dependant on specific vegetative characteristics as well grain traits. This is because yield, which represents the overall grain productivity of a plant, depends on morphological characteristics such as the degree of tillering, an architecture that allows the plant to efficiently capture light and carbon, and a stature that is resistant to lodging, the degree of panicle branching, as well as specific grain characteristics such as seed volume and shape. Moreover, there is a direct biological linkage between specific vegetative architectural traits such as tillering and plant height, and yield related traits such as panicle branching and number of seeds per panicle. The degree of branching during both vegetative and reproductive development is dependant on the development and initiation of auxiliary meristems. Several genes have been identified in this pathway and have shown to have pleiotropic effects on tillering and panicle branching (reviewed by [155]). For instance, *OsSPL14* has been shown to be an important regulator of auxiliary branching in both vegetative and reproductive stages in rice [156, 157]. Moreover, other genes such as *OsGhd8* have been reported to regulate other morphological traits such as plant height and yield through increase panicle branching [158]. The biological importance of these dependencies can also be illustrated by viewing them in the context of genetic improvement, as selection for specific architectural traits (represented by the latent variable morphology) and grain characteristics have traditionally been used as traits to improve rice productivity

in many conventional breeding programs [159, 160].

While the above example provides a plausible network structure between latent variables, edges from ionic components of salt stress to flowering time and to grain morphology are an example of instances where caution should be used to infer causation. As mentioned above, there is an inherent difference in salt tolerance and grain morphological traits between the indica and japonica subspecies. The edges observed for these two latent variables (ionic components of salt stress and grain morphology) in BN may be driven by LD between alleles associated with grain morphology and alleles for salt tolerance rather than pleitropy. Thus, given the current data set, genetic effects for grain morphology may still be conditionally dependant on ionic components of salt stress and the BN may be true, even if there is no direct overlap in the genetic mechanisms for the two traits.

We found that there are some uncertain edges among BN in Figure 6.5. For instance, direction from morphological salt response to grain morphology is supported by 65% (Tabu), 58% (Max-Min Hill Climbing), and 58% (General 2-Phase Restricted Maximization) bootstrap sampling, whereas the opposite direction is supported by 56% bootstrap sampling (Hill Climbing). An analogous uncertainty was also observed between morphology and flowering time, i.e., the path from morphology to flowering time was supported 60% (Hill Climbing), 51% (Tabu), and 52% (General 2-Phase Restricted Maximization), while the reverse direction was supported 51% (Max-Min Hill Climbing) upon bootstrapping. In addition, the two score-based algorithms captured edges between morphological salt response and flowering time with 70% and 76% bootstrapping evidence. However, this connection was not detected in the two hybrid algorithms. In general, inferring the direction of edges was harder than inferring the presence or absence of undirected edges. Finally, the whole structures of BN were evaluated in terms of the BIC score and BGe. Ranking of the networks was consistent across BIC and BGe and the two score-based algorithms produced networks with greater

goodness-of-fit than the two hybrid algorithms. The optimal network was produced by the Tabu algorithm. This is consistent with the previous study reporting that the score-based algorithm produced a better fit of networks in data on maize [55].

In conclusion, the present results show the utility of CFA and network analysis to characterize various phenotypes in rice. We showed that the joint use of BCFA and BN can be applied to predict the potential influence of external interventions or selection associated with target traits such as yield in the high-dimensional interrelated complex traits system. We contend that the approaches used herein provide greater insights than pairwise-association measures of multiple phenotypes and can be used to analyze the massive amount of diverse image-based phenomics dataset being generated by the automated plant phenomics platforms [e.g., 161]. With a large volume of complex traits being collected through phenomics, numerous opportunities to forge new research directions are generated by using network analysis for the growing number of phenotypes.

6.6 Tables

Table 6.1: Standardized factor loadings obtained from the Bayesian confirmatory factor analysis. PSD refers to the posterior standard deviation of standardized factor loadings.

Latent variable	Observed phenotype	Loading	PSD
Flowering time	Flowering time at Arkansas (Fla)	0.990	0.002
Flowering time	Flowering time at Faridpur (Flf)	0.500	0.045
Flowering time	Flowering time at Aberdeen (Flb)	0.578	0.038
Flowering time	FT ratio of Arkansas/Aberdeen (Flaa)	-0.212	0.053
Flowering time	FT ratio of Faridpur/Aberdeen (Flfa)	-0.549	0.041
Flowering time	Year07 Flowering time at Arkansas (Fla7)	0.926	0.008
Flowering time	Year06 Flowering time at Arkansas (Fla6)	0.886	0.013
Morphology	Culm habit (Cuh)	0.227	0.027
Morphology	Flag leaf length (Fll)	0.116	0.057
Morphology	Flag leaf width (Flw)	-0.044	0.058
Morphology	Plant height (Plh)	0.440	0.047
Morphology	Shoot BM Control (Sbc)	0.534	0.042
Morphology	Shoot BM Salt (Sbs)	0.456	0.048
Morphology	Root BM Control (Rbc)	0.418	0.048
Morphology	Root BM Salt (Rbs)	0.280	0.054
Morphology	Tiller No Salt (Tns)	-0.349	0.051
Morphology	Tiller No Control (Tbc)	-0.318	0.052
Morphology	Ht Lig Salt (Hls)	0.920	0.011
Morphology	Ht Lig Control (Hlc)	0.899	0.014
Morphology	Ht FE Salt (Hfs)	0.907	0.013
Morphology	Ht FE Control (Hfc)	0.925	0.011
Yield	Panicle number per plant (Pnu)	0.190	0.020
Yield	Panicle length (Pal)	0.455	0.057
Yield	Primary panicle branch number (Ppn)	0.790	0.041
Yield	Seed number per panicle (Snpp)	0.780	0.043
Yield	Panicle fertility (Paf)	-0.085	0.081
Grain Morphology	Seed length (Sl)	0.251	0.029
Grain Morphology	Seed width (Sw)	0.876	0.015
Grain Morphology	Seed volume (Sv)	0.990	0.002
Grain Morphology	Seed surface area (Ssa)	0.901	0.012
Grain Morphology	Brown rice seed length (Bsl)	0.158	0.055
Grain Morphology	Brown rice seed width (Bsw)	0.837	0.019
Grain Morphology	Brown rice surface area (Bsa)	0.902	0.012
Grain Morphology	Brown rice volume (Bvl)	0.986	0.002
Grain Morphology	Seed length/width ratio (Slwr)	-0.476	0.045
Grain Morphology	Brown rice length/width ratio (Blwr)	-0.432	0.047
Grain Morphology	Grain length McCouch2016 (Glmc)	0.047	0.064
Ionic components of salt stress	Na K Shoot (Ks)	0.983	0.003
Ionic components of salt stress	Na Shoot (Nas)	0.975	0.004
Ionic components of salt stress	K Shoot Salt (Kss)	-0.265	0.051
Ionic components of salt stress	Na K Root (Kr)	0.061	0.052
Ionic components of salt stress	Na Root (Nar)	0.001	0.053
Ionic components of salt stress	K Root Salt (Krs)	-0.095	0.052
Morphological salt response	Shoot BM Ratio (Sbr)	0.410	0.047
Morphological salt response	Root BM Ratio (Rbr)	0.395	0.051
Morphological salt response	Tiller No Ratio (Tbr)	-0.022	0.057
Morphological salt response	Ht Lig Ratio (Hlr)	0.665	0.036
Morphological salt response	Ht FE Ratio (Hfr)	0.939	0.019

6.7 Figures

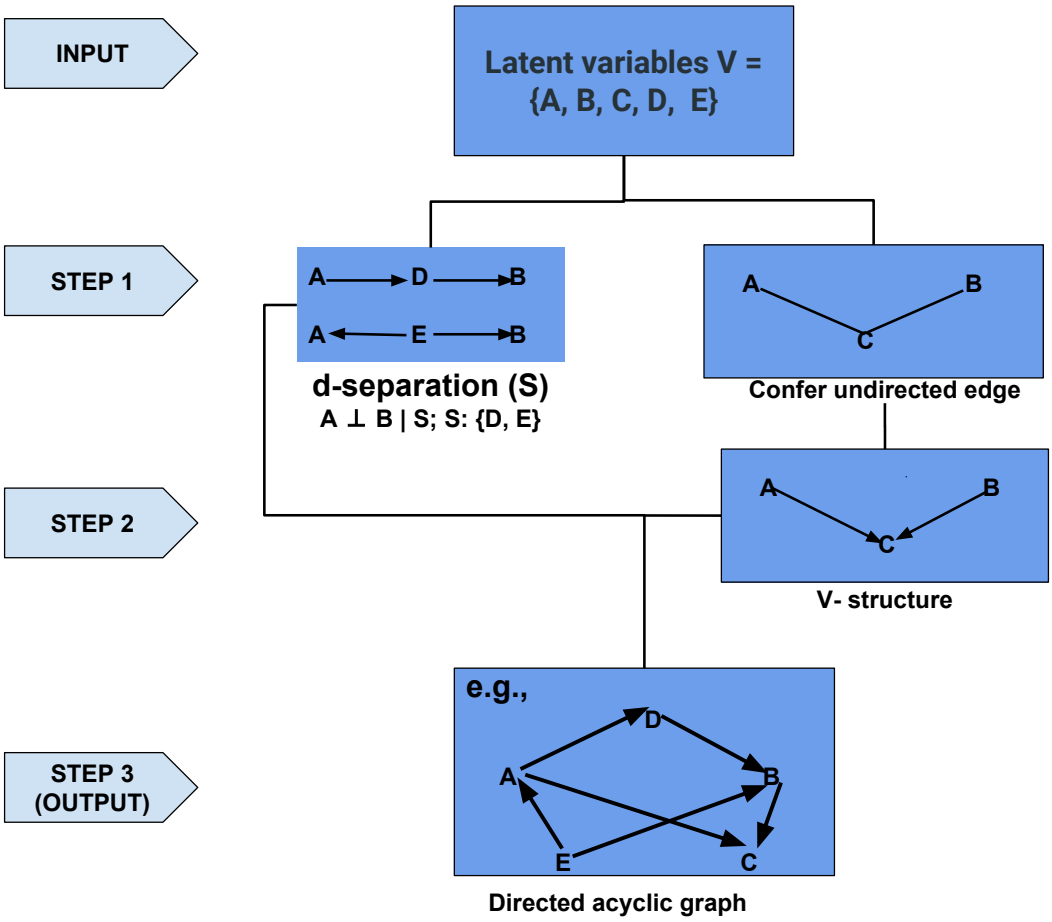

Figure 6.1: Flow diagram to illustrate the concept of constraint-based structure learning algorithm for a Bayesian network. The A, B, C, D, and E represent five nodes or latent variables. S refers to a set of d-separation. The directed acyclic graph shown in Step 3 is one possible completed partially directed acyclic graph.

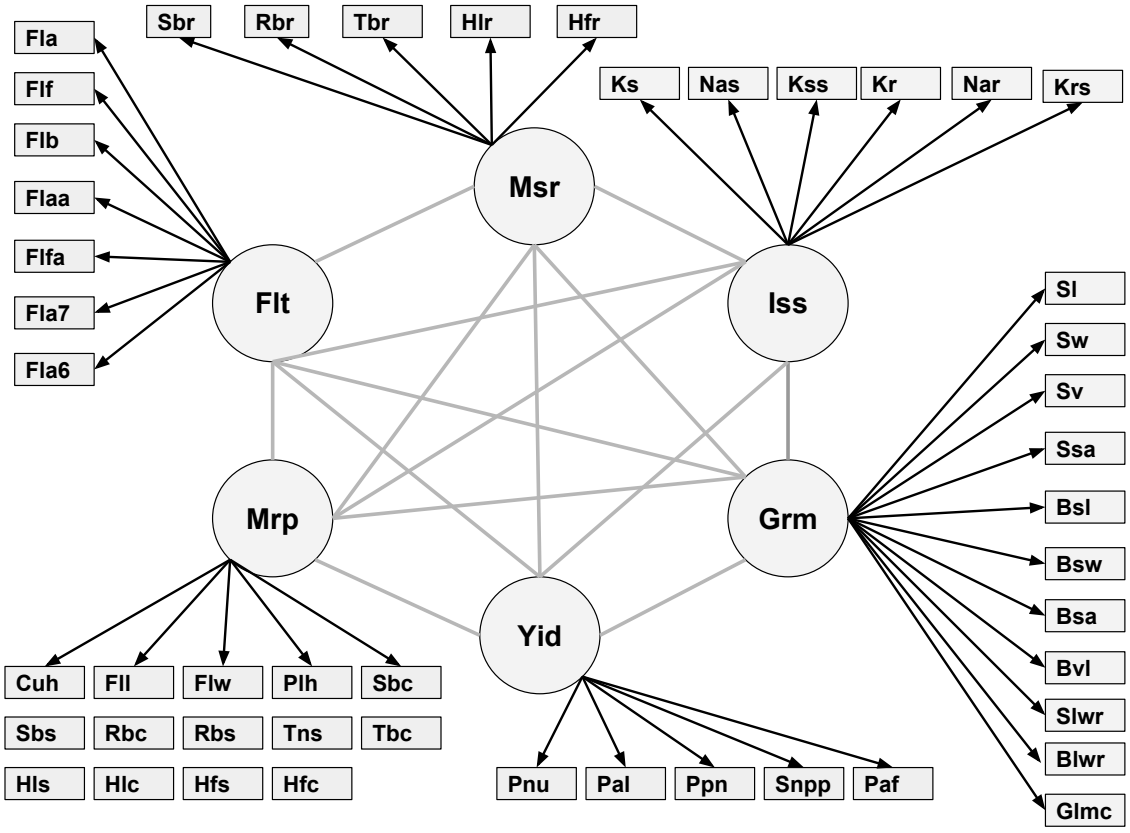

Figure 6.2: Relationship between six latent variables and observed phenotypes. Msr: morphological salt response; Iss: ionic components of salt stress; Grm: grain morphology; Yid: yield; Mrp: morphology; Flt: flowering time. Abbreviations of observed phenotypes are shown in Appendix C.

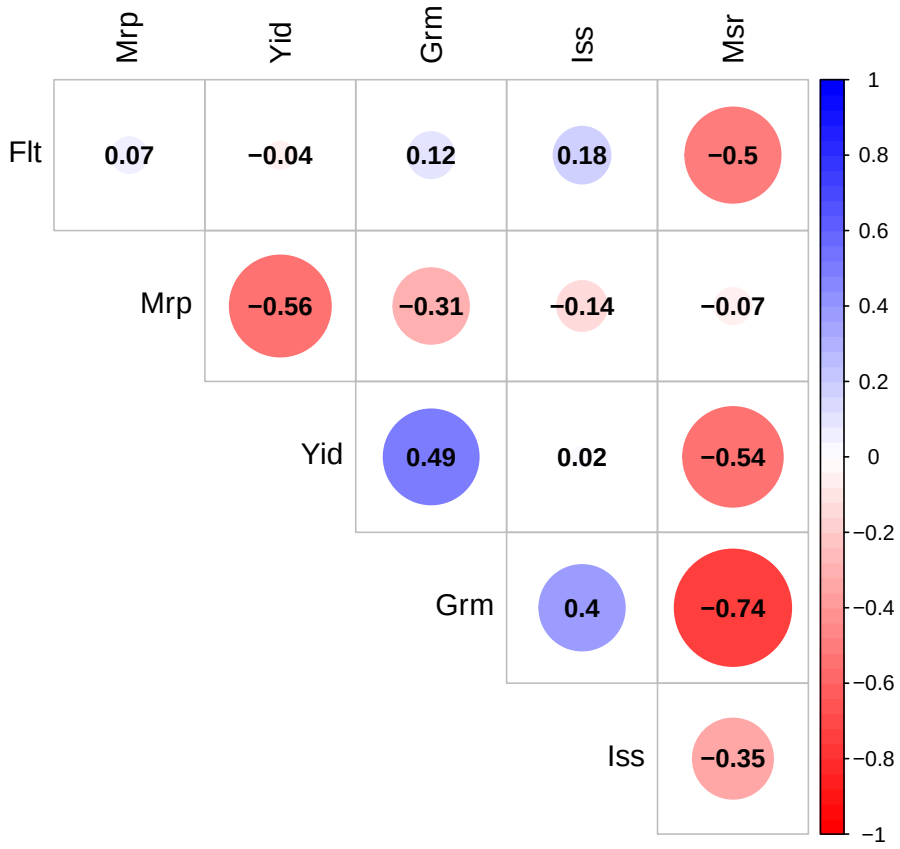

Figure 6.3: Genomic correlation of six latent variables. The size of each circle, degree of shading, and value reported correspond to the correlation between each pair of latent variables. Msr: morphological salt response; Iss: ionic components of salt stress; Grm: grain morphology; Yid: yield; Mrp: morphology; Flt: flowering time.

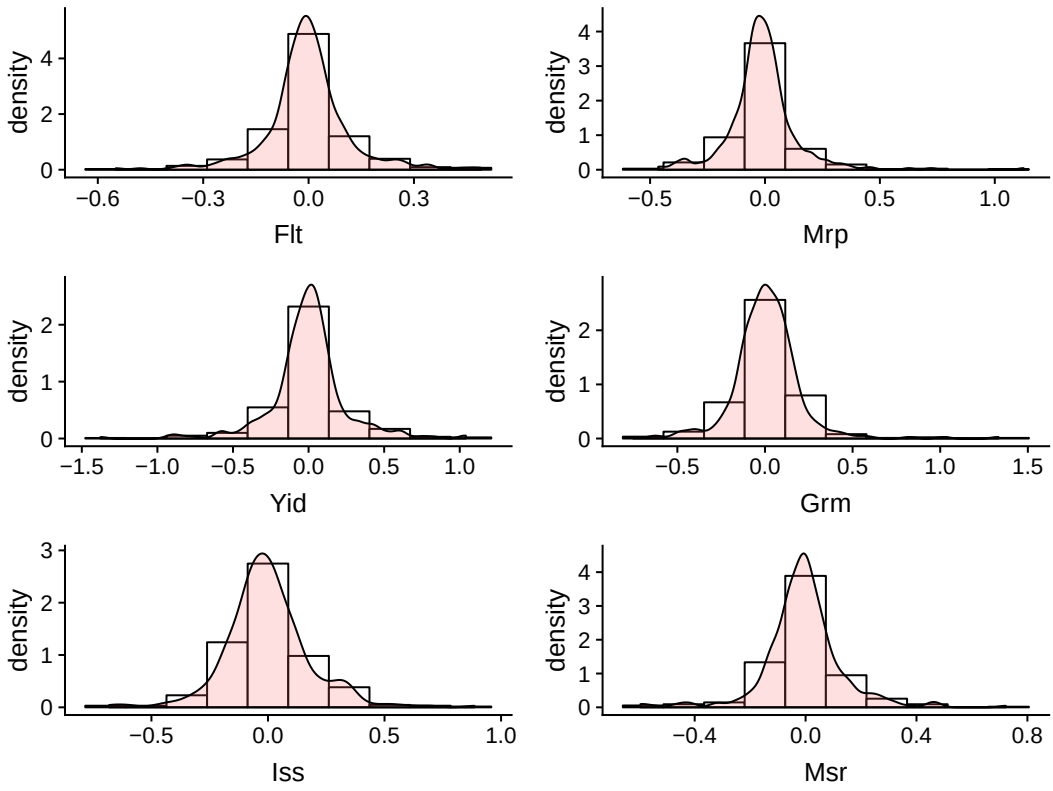

Figure 6.4: Histogram plots and density curves of six latent variables. Flt: flowering time; Mrp: morphology; Yid: yield; Grm: grain morphology; Iss: ionic components of salt stress; Msr: morphological salt response.

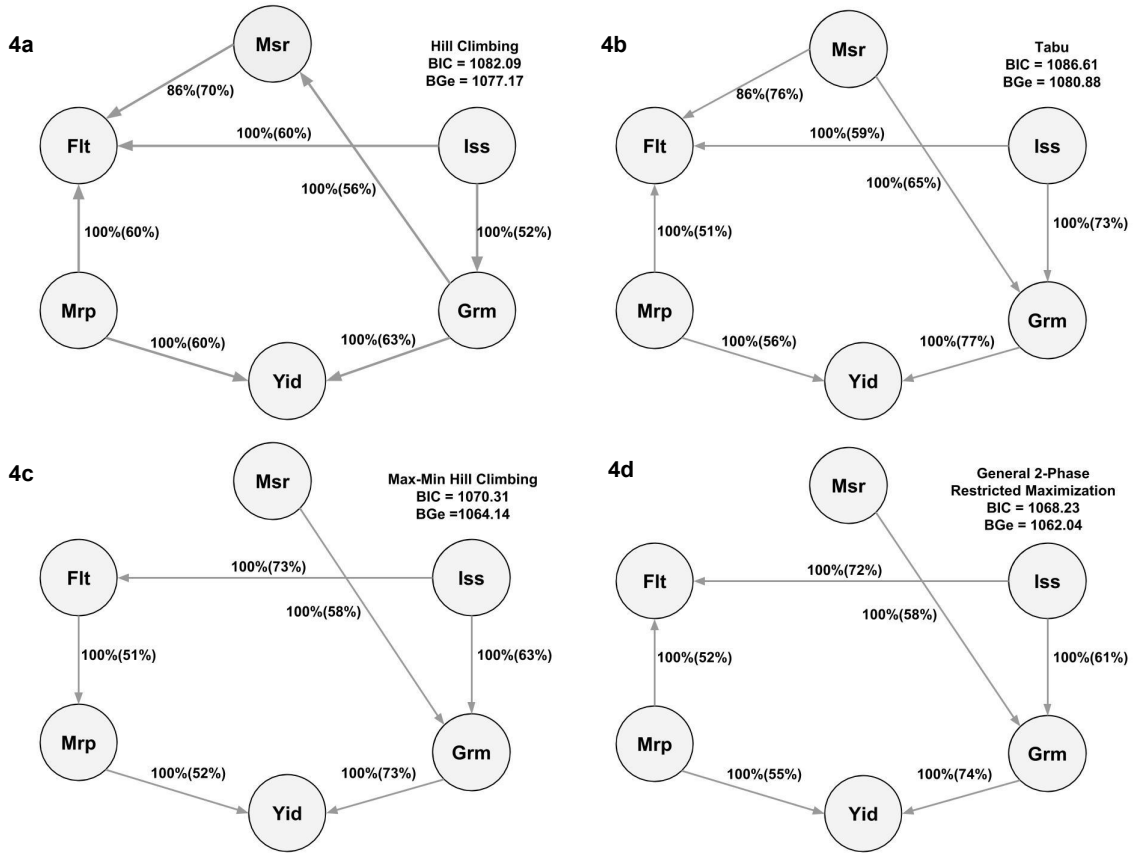

Figure 6.5: Bayesian networks between six latent variables based on two score-based (4a: Hill Climbing and 4b: Tabu) and two hybrid (4c: Max-Min Hill Climbing and 4d: General 2-Phase Restricted Maximization) algorithms. The quality of the structure was evaluated by bootstrap resampling and model averaging across 500 replications. Labels of the edges refer to the strength and direction (parenthesis) which measure the confidence of the directed edge. The strength indicates the frequency of the edge is present and the direction measures the frequency of the direction conditioned on the presence of edge. BIC: Bayesian information criterion score. BGe: Bayesian Gaussian equivalent score. Msr: morphological salt response; Iss: ionic components of salt stress; Grm: grain morphology; Yid: yield; Mrp: morphology; Flt: flowering time.

Chapter 7

Conclusions

This dissertation focuses on providing a guideline to design and model high-dimensional high-throughput phenotyping (HTP) data using whole-genome regression. Genetic connectedness measures the relatedness between individuals. A sufficient level of connectedness is required to reduce the uncertainty of ranking when comparing estimated breeding values derived from best linear unbiased prediction across different units [9]. Connectedness statistics can also be used for the choice of individuals to be phenotyped as the training data [15, 16, 18].

My connectedness studies showed that whole-genome markers can enhance the measures of connectedness compared with using pedigree information. The use of genomics enhanced the measures of connectedness and was shown to be positively correlated with cross-validation-based genome-enabled prediction accuracy. Although the rapid development of HTP technologies has significantly reduced human labor, efficient phenotyping designs for training sets are still needed. My findings and the GCA R package I developed here will be useful for researchers to study and design which individuals to be phenotyped.

The availability of HTP platforms combined with genome-wide markers has provided a suite of resources for genetic analysis of large populations. A new challenge associated with the high-dimensional data derived from HTP is how to efficiently analyze and interpret the interrelationships among phenotypes appropriately. An multi-trait model accommodates genetic or environmental covariances among traits [44, 46, 48, 49], but often hindered by a computational limitation when a large number of traits were analyzed. Under such a

circumstance, a dimensional reduction approach, including factor analysis, can be used [50, 51]. Furthermore, multi-trait model cannot detect the direction of the relationships among traits, which is usually essential for genetic selection in a complex traits system. I used a novel statistical approach by leveraging the combination of confirmatory factor analysis and Bayesian network to decipher genetic interrelationships among image-derived HTP data. The proposed statistical method will open a new opportunity to model and understand the directed relationships among high-dimensional traits. This will provide a guideline for making genetic selection decisions in complex trait systems.

References

[1] S Wright. Systems of mating. I. The biometric relations between offspring and parent. *Genetics*, 6:111–123, 1921.

[2] T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001. ISSN 0016-6731.

[3] LR Schaeffer. Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 123(4):218–223, 2006.

[4] Jessica Rutkoski, Jesse Poland, Suchismita Mondal, Enrique Autrique, Lorena González Pérez, José Crossa, Matthew Reynolds, and Ravi Singh. Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3: Genes, Genomes, Genetics*, 6(9):2799–2808, 2016.

[5] Jared Crain, Suchismita Mondal, Jessica Rutkoski, Ravi P Singh, and Jesse Poland. Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding. *The Plant Genome*, 2018.

[6] Marcus Jansen, Frank Gilmer, Bernhard Biskup, Kerstin A Nagel, Uwe Rascher, Andreas Fischbach, Sabine Briem, Georg Dreissen, Susanne Tittmann, Silvia Braun, et al. Simultaneous phenotyping of leaf growth and chlorophyll fluorescence via growscreen fluoro allows detection of stress tolerance in arabidopsis thaliana and other rosette plants. *Functional Plant Biology*, 36(11):902–914, 2009.

[7] Randy T Clark, Robert B MacCurdy, Janelle K Jung, Jon E Shaff, Susan R McCouch, Daniel J Aneshansley, and Leon V Kochian. Three-dimensional root phenotyping with a novel imaging and software platform. *Plant Physiology*, 156(2):455–465, 2011.

[8] Lingfeng Duan, Wanneng Yang, Chenglong Huang, and Qian Liu. A novel machine-vision-based facility for the automatic evaluation of yield-related traits in rice. *Plant Methods*, 7(1):44, 2011.

[9] JL Foulley, J Bouix, B Goffinet, and JM Elsen. Connectedness in genetic evaluation. In *Advances in statistical methods for genetic improvement of livestock*, pages 277–308. Springer, 1990.

[10] L A Kuehn, R M Lewis, and D R Notter. Connectedness in Targhee and Suffolk flocks participating in the United States national sheep improvement program. *Journal of Animal Science*, 87:507–515, 2009.

[11] L S Eikje and R M Lewis. Strong connectedness within Norwegian Cheviot and Fur Sheep ram circles allows reliable estimation of breeding values. *Journal of Animal Science*, 93:3322–3330, 2015.

[12] JL Foulley, LR Schaeffer, H Song, and JW Wilton. 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.

[13] Eric Hanocq, Didier Boichard, and Jean Louis Foulley. A simulation study of the effect of connectedness on genetic trend. *Genetics Selection Evolution*, 28(1):67, 1996.

[14] L A Kuehn, D R Notter, G J Nieuwhof, and R M Lewis. Changes in connectedness over time in alternative sheep sire referencing schemes. *Journal of Animal Science*, 86: 536–544, 2008.

[15] M. Pszczola, T Strabel, J A M van Arendonk, and M P L Calus. The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. *Journal of Dairy Science*, 95:5412–5421, 2012.

[16] Julio Isidro, Jean-Luc Jannink, Deniz Akdemir, Jesse Poland, Nicolas Heslot, and Mark E Sorrells. Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics*, 128(1):145–158, 2015.

[17] S Maenhout, B De Baets, and G Haesaert. Graph-based data selection for the construction of genomic prediction models. *Genetics*, 185:1463–1475, 2010.

[18] Renaud Rincent, Denis Laloë, Stéphane Nicolas, Thomas Altmann, Dominique Brunel, Pedro Revilla, Victor M Rodriguez, J Moreno-Gonzalez, A Melchinger, Eva Bauer, 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.

[19] C R Henderson. *Applications of linear models in animal breeding*. University of Guelph;, Third edition, Edited by Schaeffer LR. Guelph, 1984.

[20] B W Kennedy and D Trus. Considerations on genetic connectedness between management units under an animal model. *Journal of Animal Science*, 71:2341–2352, 1993.

[21] D Laloë. Precision and information in linear models of genetic evaluation. *Genetics Selection Evolution*, 25:557, 1993.

[22] RM Lewis, RE Crump, G Simm, and R Thompson. Assessing connectedness in across-flock genetic evaluations. *Proc. Brit. Soc. Anim. Sci*, 121, 1999.

[23] John B Holmes, Ken G Dodds, and Michael A Lee. Estimation of genetic connectedness

diagnostics based on prediction errors without the prediction error variance–covariance matrix. *Genetics Selection Evolution*, 49(1):29, 2017.

[24] Audrey Darrigues, Jack Hall, Esther van der Knaap, David M Francis, Nancy Dujmovic, and Simon Gray. Tomato analyzer-color test: a new tool for efficient digital phenotyping. *Journal of the American Society for Horticultural Science*, 133(4):579–586, 2008.

[25] Antonio Girolami, Fabio Napolitano, Daniela Faraone, and Ada Braghieri. Measurement of meat color using a computer vision system. *Meat Science*, 93(1):111–118, 2013.

[26] Abhinav Kumar, Sonal Saxena, Sameer Shrivastava, Vandana Bharti, Upendra Kumar, and Kuldeep Dhama. Hyperspectral imaging (hsi): Applications in animal and dairy sector. *Journal of Experimental Biology and Agricultural Sciences*, 4(4):448–461, 2016.

[27] Cristina González-Flor, Lydia Serrano, Gil Gorchs, and Josep M Pons. Assessment of grape yield and composition using reflectance-based indices in rainfed vineyards. *Agronomy Journal*, 106(4):1309–1316, 2014.

[28] JRR Dórea, GJM Rosa, KA Weld, and LE Armentano. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. *Journal of Dairy Science*, 101(7):5878–5889, 2018.

[29] María L Pérez-Bueno, Mónica Pineda, Francisco M Cabeza, and Matilde Barón. Multicolor fluorescence imaging as a candidate for disease detection in plant phenotyping. *Frontiers in Plant Science*, 7:1790, 2016.

[30] Charlotte Møller Andersen and Grith Mortensen. Fluorescence spectroscopy: A rapid

tool for analyzing dairy products. *Journal of Agricultural and Food Chemistry*, 56(3): 720–729, 2008.

[31] JM Blonquist Jr, J M Norman, and Bruce Bugbee. Automated measurement of canopy stomatal conductance based on infrared temperature. *Agricultural and Forest Meteorology*, 149(11):1931–1945, 2009.

[32] JF Hurnik, S De Boer, and AB Webster. Detection of health disorders in dairy cattle utilizing a thermal infrared scanning technique. *Canadian Journal of Animal Science*, 64(4):1071–1073, 1984.

[33] Joseph A Loughmiller, Mark F Spire, Steve S Dritz, Bradley W Fenwick, Mohammad H Hosni, and Steven B Hogge. Relationship between mean body surface temperature measured by use of infrared thermography and ambient temperature in clinically normal pigs and pigs inoculated with actinobacillus pleuropneumoniae. *American Journal of Veterinary Research*, 62(5):676–681, 2001.

[34] Latif Emrah Yanmaz, Zafer Okumus, and Elif Dogan. Instrumentation of thermography and its applications in horses. *Journal of Animal and Veterinary Advances*, 6(7): 858–62, 2007.

[35] Stefan Paulus, Jan Dupuis, Anne-Katrin Mahlein, and Heiner Kuhlmann. Surface feature based classification of plant organs from 3d laserscanned point clouds for plant phenotyping. *BMC Bioinformatics*, 14(1):238, 2013.

[36] Mirwaes Wahabzada, Stefan Paulus, Kristian Kersting, and Anne-Katrin Mahlein. Automated interpretation of 3d laserscanned point clouds for plant organ segmentation. *BMC Bioinformatics*, 16(1):248, 2015.

[37] Ulrich Weiss and Peter Biber. Plant detection and mapping for agricultural robots using a 3d lidar sensor. *Robotics and Autonomous Systems*, 59(5):265–273, 2011.

[38] William D Simonson, Harriet D Allen, and David A Coomes. Applications of airborne lidar for the assessment of animal species diversity. *Methods in Ecology and Evolution*, 5(8):719–729, 2014.

[39] Ryan F McCormick, Sandra K Truong, and John E Mullet. 3d sorghum reconstructions from depth images identify qtl regulating shoot architecture. *Plant Physiology*, 172(2): 823–834, 2016.

[40] Chunlei Xia, Longtan Wang, Bu-Keun Chung, and Jang-Myung Lee. In situ 3d segmentation of individual plant leaves using a rgb-d camera for agricultural automation. *Sensors*, 15(8):20463–20479, 2015.

[41] Qiming Zhu, Jinchang Ren, David Barclay, Samuel McCormack, and Willie Thomson. Automatic animal detection from kinect sensed images for livestock monitoring and assessment. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pages 1154–1157. IEEE, 2015.

[42] Arthur FA Fernandes, João RR Dórea, Robert Fitzgerald, William Herring, and Guilherme JM Rosa. A novel automated system to acquire biometric and morphological measurements and predict body weight of pigs via 3d computer vision. *Journal of Animal Science*, 97(1):496–508, 2019.

[43] A Cominotte, AFA Fernandes, JRR Dorea, GJM Rosa, MM Ladeira, EHCB van Cleef, GL Pereira, WA Baldassini, and OR Machado Neto. Automated computer vision

system to predict body weight and average daily gain in beef cattle during growing and finishing phases. *Livestock Science*, 232:103904, 2020.

[44] CR Henderson and RL Quaas. Multiple trait evaluation using relatives' records. *Journal of Animal Science*, 43(6):1188–1197, 1976.

[45] Raphael A Mrode. *Linear models for the prediction of animal breeding values*. Cabi, 2014.

[46] Yi Jia and Jean-Luc Jannink. Multiple trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, pages genetics–112, 2012.

[47] Mario PL Calus and Roel F Veerkamp. Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution*, 43(1):26, 2011.

[48] Diego Jarquín, José Crossa, Xavier Lacaze, Philippe Du Cheyron, Joëlle Daucourt, Josiane Lorgeou, François Piraux, Laurent Guerreiro, Paulino Pérez, Mario Calus, et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, 127(3):595–607, 2014.

[49] Marco Lopez-Cruz, Jose Crossa, David Bonnett, Susanne Dreisigacker, Jesse Poland, Jean-Luc Jannink, Ravi P Singh, Enrique Autrique, and Gustavo de los Campos. Increased prediction accuracy in wheat breeding trials using a marker× environment interaction genomic selection model. *G3: Genes, Genomes, Genetics*, 5(4):569–582, 2015.

[50] Gustavo de los Campos and Daniel Gianola. Factor analysis models for structuring covariance matrices of additive genetic effects: a bayesian implementation. *Genetics Selection Evolution*, 39(5):481, 2007.

[51] F Peñagaricano, BD Valente, JP Steibel, RO Bates, CW Ernst, H Khatib, and GJM Rosa. Searching for causal networks involving latent variables in complex traits: application to growth, carcass, and meat quality traits in pigs. *Journal of Animal Science*, 93(10):4617–4623, 2015.

[52] Richard E Neapolitan et al. *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[53] G Morota, BD Valente, GJM Rosa, KA Weigel, and D Gianola. An assessment of linkage disequilibrium in holstein cattle using a bayesian network. *Journal of Animal Breeding and Genetics*, 129(6):474–487, 2012.

[54] Alencar Xavier, Benjamin Hall, Shaun Casteel, William Muir, and Katy Martin Rainey. Using unsupervised learning techniques to assess interactions among complex traits in soybeans. *Euphytica*, 213(8):200, 2017.

[55] Katrin Töpner, Guilherme JM Rosa, Daniel Gianola, and Chris-Carolin Schön. Bayesian networks illustrate genomic and residual trait connections in maize (Zea mays L.). *G3: Genes, Genomes, Genetics*, 7(8):2779–2789, 2017.

[56] Larry A Kuehn, Ronald M Lewis, and David R. Notter. Managing the risk of comparing estimated breeding values across flocks or herds through connectedness: a review and application. *Genetics Selection Evolution*, 39:225, 2007.

[57] R Rincent, D Laloë, S Nicolas, T Altmann, D Brunel, P Revilla, V M Rodríguez, J Moreno-Gonzalez, A Melchinger, E Bauer, C C Schoen, N Meyer, C Giauffret, C Bauland, P Jamin, J Laborde, H Monod, P Flament, A Charcosset, and L Moreau. 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*, 192:715–728, 2012.

[58] Julio Isidro, Jean-Luc Jannink, Deniz Akdemir, Jesse Poland, Nicolas Heslot, and Mark E Sorrells. Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics*, 128:145, 2015.

[59] W G Hill and B S Weir. Variation in actual relationship as a consequence of mendelian sampling and linkage. *Genetics Research*, 93:47–64, 2011.

[60] William Valdar, Leah C Solberg, Dominique Gauguier, Stephanie Burnett, Paul Klenerman, William O Cookson, Martin S Taylor, J Nicholas P Rawlins, Richard Mott, and Jonathan Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, 38:879–887, 2006.

[61] Leah C Solberg, William Valdar, Dominique Gauguier, Graciela Nunez, Amy Taylor, Stephanie Burnett, Carmen Arboledas-Hita, Polinka Hernandez-Pliego, Stuart Davidson, Peter Burns, Shoumo Bhattacharya, Tertius Hough, Douglas Higgs, Paul Klenerman, William O Cookson, Youming Zhang, Robert M Deacon, J Nicholas P Rawlins, Richard Mott, and Jonathan Flint. A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice. *Mamm Genome*, 17:129–146, 2006.

[62] Andrés Legarra, Christéle Robert-Granié, Eduardo Manfredi, and Jean-Michel Elsen. Performance of genomic selection in mice. *Genetics*, 180:611–618, 2008.

[63] Valentin Wimmer, Theresa Albrecht, Hans-Juergen Auinger, Chris-Carolin Schoen with contributions by Malena Erbe, Ulrike Ober, and Christian Reimer. *synbreedData: Data for the Synbreed package*, 2015.

[64] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, New York, 1990.

[65] A P Reynolds, G Richards, B de la Iglesia, and V J Rayward-Smith. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5:475, 2006.

[66] R M Lewis, R E Crump, G Simm, and R Thompson. Assessing connectedness in across-flock genetic evaluations. In *Proc. Br. Soc. Anim. Sci.*, page 121, Scarborough, UK., 1999.

[67] D Laloë, F Phocas, and F Ménissier. Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. *Genetics Selection Evolution*, 28:359, 1996.

[68] S Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56: 330–338, 1922.

[69] P M VanRaden. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91:4414–4423, 2008.

[70] Migue Toro, Carmen Barragán, Cristina Óvilo, Jaime Rodrigañez, Carmen Rodriguez, and Luis Siliá. Estimation of coancestry in Iberian pigs using molecular markers. *Conservation Genetics*, 3:309–320, 2002.

[71] Joseph E Powell, Peter M Visscher, and Michael E Goddard. Reconciling the analysis of ibd and ibs in complex trait studies. *Nature Reviews Genetics*, 11:800–805, 2010.

[72] P M VanRaden. Genomic measures of relationship and inbreeding. *Interbull Bulletin*, 37:33–36, 2007.

[73] M A Toro, L A García-Cortés, and A Legarra. A note on the rationale for estimating genealogical coancestry from molecular markers. *Genetics Selection Evolution*, 43:27, 2011.

[74] Z G Vitezica, Ignacio Aguilar, Ignacy Misztal, and A Legarra. Bias in genomic predictions for populations under selection. *Genetics Research*, 93:357–366, 2011.

[75] M Momen, A A Mehrgardi, A Sheikhy, A Esmailizadeh, M A Fozi, A Kranis, B D Valente, G J Rosa, and D Gianola. A predictive assessment of genetic correlations between traits in chickens using markers. *Genetics Selection Evolution*, 49:16, 2017.

[76] A Legarra, I Aguilar, and I Misztal. A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, 92:4656–4663, 2009.

[77] O F Christensen and M S Lund. Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution*, 42:2, 2010.

[78] J L Foulley, E Hanocq, and D Boichard. A criterion for measuring the degree of connectedness in linear models of genetic evaluation. *Genetics Selection Evolution*, 24: 315–330, 1992.

[79] M N Fouilloux, V Clément, and D Laloë. Measuring connectedness among herds in mixed linear models: from theory to practice in large-sized genetic evaluations. *Genetics Selection Evolution*, 40:145–159, 2008.

[80] R A Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[81] M Goddard. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136:245–257, 2009.

[82] R Fernando and M Grossman. Marker assisted selection using best linear unbiased prediction. *Genetics Selection Evolution*, 21:467–477, 1989.

[83] H Yu, ML Spangler, RM Lewis, and G Morota. Genomic relatedness strengthens genetic connectedness across management units. *G3: Genes, Genomes, Genetics*, 7: 543–3556, 2017.

[84] M Sargolzaei and FS Schenkel. QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, 25:680–681, 2009.

[85] G Morota and D Gianola. Kernel-based whole-genome prediction of complex traits: a review. *Frontiers in Genetics*, 5, 2014.

[86] Paulino Pérez and Gustavo de los Campos. Genome-wide regression and prediction with the bglr statistical package. *Genetics*, 198(2):483–495, 2014. ISSN 0016-6731. doi: 10.1534/genetics.114.164442.

[87] DL Weeks and DR Williams. A note on the determination of connectedness in an n-way cross classification. *Technometrics*, 6(3):319–324, 1964.

[88] JA Eccleston and A Hedayat. On the theory of connected designs: characterization and optimality. *The Annals of Statistics*, pages 1238–1255, 1974.

[89] SR Searle. *Linear models*. John Wiley & Sons, New York, 1986.

[90] SR Miraei Ashtiani and JW James. Efficient use of link rams in merino sire reference schemes. In *Proc 9th Conf. Aust. Assoc. Anim. Breed. Genet*, pages 24–27, 1991.

[91] Eric Hanocq and Didier Boichard. Connectedness in the french holstein cattle population. *Genetics Selection Evolution*, 31(2):163, 1999.

[92] Ulrike Ober, Julien F Ayroles, Eric A Stone, Stephen Richards, Dianhui Zhu, Richard A Gibbs, Christian Stricker, Daniel Gianola, Martin Schlather, Trudy FC Mackay, et al. Using whole-genome sequence data to predict quantitative trait phenotypes in drosophila melanogaster. *PLoS Genetics*, 8(5):e1002685, 2012.

[93] Hans D Daetwyler, Ricardo Pong-Wong, Beatriz Villanueva, and John A Woolliams. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, 185 (3):1021–1031, 2010.

[94] Gota Morota. Shinygpas: interactive genomic prediction accuracy simulator based on deterministic formulas. *Genetics Selection Evolution*, 49(1):91, 2017.

[95] PH Petersen. A test for connectedness fitted for the two-way blup-sire evaluation. *Acta Agriculturae Scandinavica*, 28(4):360–362, 1978.

[96] RL Fernando, D Gianola, and M Grossman. Identifying all connected subsets in a two-way classification without interaction. *Journal of Dairy Science*, 66(6):1399–1402, 1983.

[97] JL Foulley, J Bouix, B Goffinet, et al. Connectedness in genetic evaluation. In *Advances in Statistical Methods for Genetic Improvement of Livestock*, pages 277–308. Springer, 1990.

[98] LA Kuehn, DR Notter, GJ Nieuwhof, and RM Lewis. Changes in connectedness over time in alternative sheep sire referencing schemes. *Journal of Animal Science*, 86(3): 536–544, 2008.

[99] Haipeng Yu, Matthew L Spangler, Ronald M Lewis, and Gota Morota. Genomic relatedness strengthens genetic connectedness across management units. *G3: Genes, Genomes, Genetics*, 7(10):3543–3556, 2017.

[100] Haipeng Yu, Matthew L Spangler, Ronald M Lewis, and Gota Morota. Do stronger measures of genomic connectedness enhance prediction accuracies across management units? *Journal of Animal Science*, 96(11):4490–4500, 2018.

[101] Mehdi Momen and Gota Morota. Quantifying genomic connectedness and prediction accuracy from additive and non-additive gene actions. *Genetics Selection Evolution*, 50(1):45, 2018.

[102] BW Kennedy and D Trus. Considerations on genetic connectedness between management units under an animal model. *Journal of Animal Science*, 71(9):2341–2352, 1993.

[103] Denis Laloë, Florence Phocas, and Francois Menissier. Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. *Genetics Selection Evolution*, 28(4):359, 1996.

[104] PK Mathur, BP Sullivan, and JP Chesnais. Measuring connectedness: concept and application to a large industry breeding program. In *Proc. 7th World Congr. Genet. Appl. to Livest. Prod*, volume 19, page 23, 2002.

[105] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

[106] Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011. doi: 10.18637/jss.v040.i08. URL http://www.jstatsoft.org/v40/i08/.

[107] Hadley Wickham and Winston Chang. Devtools: Tools to make developing r packages easier. *R package version*, 1(0):9000, 2016.

[108] PJ Kaufman, L & Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, New York, 1990.

[109] Paul M VanRaden. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423, 2008.

[110] Nadia Shakoor, Scott Lee, and Todd C Mockler. High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Current Opinion in Plant Biology*, 38:184–192, 2017.

[111] José Luis Araus and Jill E Cairns. Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science*, 19(1):52–61, 2014.

[112] José Luis Araus, Shawn C Kefauver, Mainassara Zaman-Allah, Mike S Olsen, and Jill E Cairns. Translating high-throughput phenotyping into genetic gain. *Trends in Plant Science*, 2018.

[113] Llorenç Cabrera-Bosquet, Christian Fournier, Nicolas Brichet, Claude Welcker, Benoît Suard, and François Tardieu. High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. *New Phytologist*, 212(1):269–281, 2016.

[114] Jin Sun, Jessica E Rutkoski, Jesse A Poland, José Crossa, Jean-Luc Jannink, and Mark E Sorrells. Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. *The Plant Genome*, 2017.

[115] Yonghong Wang and Jiayang Li. Genes controlling plant architecture. *Current Opinion in Biotechnology*, 17(2):123–129, 2006.

[116] Shuen-Fang Lo, Show-Ya Yang, Ku-Ting Chen, Yue-Ie Hsing, Jan AD Zeevaart, Liang-Jwu Chen, and Su-May Yu. A novel class of gibberellin 2-oxidases control semid-

warfism, tillering, and root development in rice. *The Plant Cell*, 20(10):2603–2618, 2008.

[117] Mikihisa Umehara, Atsushi Hanada, Satoko Yoshida, Kohki Akiyama, Tomotsugu Arite, Noriko Takeda-Kamiya, Hiroshi Magome, Yuji Kamiya, Ken Shirasu, Koichi Yoneyama, et al. Inhibition of shoot branching by new terpenoid plant hormones. *Nature*, 455(7210):195, 2008.

[118] Anjanabha Bhattacharya, Sofia Kourmpetli, and Michael R Davey. Practical applications of manipulating plant architecture by regulating gibberellin metabolism. *Journal of Plant Growth Regulation*, 29(2):249–256, 2010.

[119] Philip B Brewer, Hinanit Koltai, and Christine A Beveridge. Diverse roles of strigolactones in plant development. *Molecular Plant*, 6(1):18–28, 2013.

[120] Feng Zhou, Qibing Lin, Lihong Zhu, Yulong Ren, Kunneng Zhou, Nitzan Shabek, Fuqing Wu, Haibin Mao, Wei Dong, Lu Gan, et al. D14–scf d3-dependent degradation of d53 regulates strigolactone signalling. *Nature*, 504(7480):406, 2013.

[121] Karl G Jöreskog. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202, 1969.

[122] Keyan Zhao, Chih-Wei Tung, Georgia C Eizenga, Mark H Wright, M Liakat Ali, Adam H Price, Gareth J Norton, M Rafiqul Islam, Andy Reynolds, Jason Mezey, et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in oryza sativa. *Nature Communications*, 2:467, 2011.

[123] Susan R McCouch, Mark H Wright, Chih-Wei Tung, Lyza G Maron, Kenneth L McNally, Melissa Fitzgerald, Namrata Singh, Genevieve DeClerck, Francisco Agosto-

Perez, Pavel Korniliev, et al. Open access resources for genome-wide association mapping in rice. *Nature Communications*, 7:10532, 2016.

[124] Malachy T Campbell, Nonoy Bandillo, Fouad Razzaq A Al Shiblawi, Sandeep Sharma, Kan Liu, Qian Du, Aaron J Schmitz, Chi Zhang, Anne-Aliénor Véry, Aaron J Lorenz, et al. Allelic variants of oshkt1; 1 underlie the divergence between indica and japonica subspecies of rice (oryza sativa) for root sodium content. *PLoS Genetics*, 13(6): e1006823, 2017.

[125] George Acquaah. *Principles of plant genetics and breeding*. John Wiley & Sons, 2009.

[126] Timothy A Brown. *Confirmatory factor analysis for applied research*. Guilford Publications, 2014.

[127] Edgar Merkle and Yves Rosseel. blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software, Articles*, 85(4):1–30, 2018. ISSN 1548-7660. doi: 10.18637/jss.v085.i04. URL https://www.jstatsoft.org/v085/i04.

[128] Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 125.10. Vienna, Austria., 2003.

[129] Matthew Denwood. runjags: An r package providing interface utilities, model templates, parallel computing methods and additional distributions for mcmc models in jags. *Journal of Statistical Software, Articles*, 71(9):1–25, 2016. ISSN 1548-7660. doi: 10.18637/jss.v071.i09. URL https://www.jstatsoft.org/v071/i09.

[130] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

[131] Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

[132] Sik-Yum Lee and Xin-Yuan Song. *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. John Wiley & Sons, 2012.

[133] Terrance P Callanan and David A Harville. *Some new algorithms for computing maximum likelihood estimates of variance components*. Iowa State University. Department of Statistics. Statistical Laboratory, 1989.

[134] AI Vazquez, DM Bates, GJM Rosa, D Gianola, and KA Weigel. An r package for fitting generalized linear mixed models in animal breeding 1. *Journal of Animal Science*, 88 (2):497–504, 2010.

[135] Marco Scutari and Jean-Baptiste Denis. *Bayesian networks: with examples in R*. Chapman and Hall/CRC, 2014.

[136] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pages 255–270, New York, NY, USA, 1991. Elsevier Science Inc. ISBN 0-444-89264-8. URL http://dl.acm.org/citation.cfm?id=647233.719736.

[137] Marco Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software, Articles*, 35(3):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v035.i03. URL https://www.jstatsoft.org/v035/i03.

[138] John M Hickey, Tinashe Chiurugwi, Ian Mackay, Wayne Powell, Andre Eggen, Andrzej Kilian, Chris Jones, Claudia Canales, Dario Grattapaglia, Filippo Bassi, et al. Genomic

prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics*, 49(9):1297, 2017.

[139] Gota Morota and Daniel Gianola. Evaluation of linkage disequilibrium in wheat with an l1-regularized sparse markov network. *Theoretical and Applied Genetics*, 126(8): 1991–2002, 2013.

[140] Bruno D Valente, Gota Morota, Francisco Peñagaricano, Daniel Gianola, Kent Weigel, and Guilherme JM Rosa. The causal meaning of genomic predictors and how it affects construction and comparison of genome-enabled selection models. *Genetics*, 200(2): 483–494, 2015.

[141] Daniel Gianola, Gustavo de los Campos, Miguel A Toro, Hugo Naya, Chris-Carolin Schön, and Daniel Sorensen. Do molecular markers inform about pleiotropy? *Genetics*, pages genetics–115, 2015.

[142] Guilherme JM Rosa, Bruno D Valente, Gustavo de los Campos, Xiao-Lin Wu, Daniel Gianola, and Martinho A Silva. Inferring causal phenotype networks using structural equation models. *Genetics Selection Evolution*, 43(1):6, 2011.

[143] Xuehui Huang, Tao Sang, Qiang Zhao, Qi Feng, Yan Zhao, Canyang Li, Chuanrang Zhu, Tingting Lu, Zhiwu Zhang, Meng Li, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, 42(11):961, 2010.

[144] Georgia C Eizenga, Md Ali, Rolfe J Bryant, Kathleen M Yeater, Anna M McClung, Susan R McCouch, et al. Registration of the rice diversity panel 1 for genomewide association studies. *Journal of Plant Registrations*, 8(1):109–116, 2014.

[145] Michael J Thomson, Abdelbagi M Ismail, Susan R McCouch, and David J Mackill.

Marker assisted breeding. In *Abiotic Stress Adaptation in Plants*, pages 451–469. Springer, 2009.

[146] Michael J Thomson, Marjorie de Ocampo, James Egdane, M Akhlasur Rahman, Andres Godwin Sajise, Dante L Adorada, Ellen Tumimbang-Raiz, Eduardo Blumwald, Zeba I Seraj, Rakesh K Singh, et al. Characterizing the saltol quantitative trait locus for salinity tolerance in rice. *Rice*, 3(2-3):148–160, 2010.

[147] Rana Munns and Mark Tester. Mechanisms of salinity tolerance. *Annual Review of Plant Biology*, 59:651–681, 2008.

[148] Zhong-Hai Ren, Ji-Ping Gao, Le-Gong Li, Xiu-Ling Cai, Wei Huang, Dai-Yin Chao, Mei-Zhen Zhu, Zong-Yang Wang, Sheng Luan, and Hong-Xuan Lin. A rice quantitative trait locus for salt tolerance encodes a sodium transporter. *Nature Genetics*, 37(10): 1141, 2005.

[149] Caitlin S Byrt, J Damien Platten, Wolfgang Spielmeyer, Richard A James, Evans S Lagudah, Elizabeth S Dennis, Mark Tester, and Rana Munns. Hkt1; 5-like cation transporters linked to na+ exclusion loci in wheat, nax2 and kna1. *Plant Physiology*, 143(4):1918–1928, 2007.

[150] Tomoaki Horie, Felix Hauser, and Julian I Schroeder. Hkt transporter-mediated salinity resistance mechanisms in arabidopsis and monocot crop plants. *Trends in Plant Science*, 14(12):660–668, 2009.

[151] Rana Munns, Richard A James, Bo Xu, Asmini Athman, Simon J Conn, Charlotte Jordans, Caitlin S Byrt, Ray A Hare, Stephen D Tyerman, Mark Tester, et al. Wheat grain yield on saline soils is improved by an ancestral na+ transporter gene. *Nature Biotechnology*, 30(4):360, 2012.

[152] RKM Hay. Harvest index: a review of its use in plant breeding and crop physiology. *Annals of Applied Biology*, 126(1):197–216, 1995.

[153] Shaobing Peng, Gurdev S Khush, Parminder Virk, Qiyuan Tang, and Yingbin Zou. Progress in ideotype breeding to increase rice yield potential. *Field Crops Research*, 108(1):32–38, 2008.

[154] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 052189560X, 9780521895606.

[155] Wei-hong Liang, Fei Shang, Qun-ting Lin, Chen Lou, and Jing Zhang. Tillering and panicle branching genes in rice. *Gene*, 537(1):1–5, 2014.

[156] Yongqing Jiao, Yonghong Wang, Dawei Xue, Jing Wang, Meixian Yan, Guifu Liu, Guojun Dong, Dali Zeng, Zefu Lu, Xudong Zhu, et al. Regulation of osspl14 by osmir156 defines ideal plant architecture in rice. *Nature Genetics*, 42(6):541, 2010.

[157] Kotaro Miura, Mayuko Ikeda, Atsushi Matsubara, Xian-Jun Song, Midori Ito, Kenji Asano, Makoto Matsuoka, Hidemi Kitano, and Motoyuki Ashikari. Osspl14 promotes panicle branching and higher grain productivity in rice. *Nature Genetics*, 42(6):545, 2010.

[158] Wen-Hao Yan, Peng Wang, Hua-Xia Chen, Hong-Ju Zhou, Qiu-Ping Li, Chong-Rong Wang, Ze-Hong Ding, Yu-Shan Zhang, Si-Bin Yu, Yong-Zhong Xing, et al. A major qtl, ghd8, plays pleiotropic roles in regulating grain productivity, plant height, and heading date in rice. *Molecular Plant*, 4(2):319–330, 2011.

[159] ED Redona and DJ Mackill. Quantitative trait locus analysis for rice panicle and grain characteristics. *Theoretical and Applied Genetics*, 96(6-7):957–963, 1998.

[160] Rongyu Huang, Liangrong Jiang, Jingsheng Zheng, Tiansheng Wang, Houcong Wang, Yumin Huang, and Zonglie Hong. Genetic bases of rice grain shape: so many genes, so little known. *Trends in Plant Science*, 18(4):218–226, 2013.

[161] R T Furbank and M Tester. Phenomics-technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, 16:635–644, 2011.

[162] Jian Yang, Sang Hong Lee, Michael E Goddard, and Peter M Visscher. Genome-wide complex trait analysis (gcta): methods, data analyses, and interpretations. pages 215–236, 2013.

Appendices

Appendix A

Prediction error correlation statistic across units

When a population is divided into two management units, and relatedness between those two units is based on the **G** matrix, the flock connectedness or unit connectedness correlation (r) of Kuehn et al. [14] always yields an estimate of -1. Here, we wish to illustrate that result. Flock connectedness is derived by averaging the relevant components of PEC and PEV followed by taking their ratio. Suppose that there are two units and the numbers of individuals in units i' and j' are $n_{i'}$ and $n_{j'}$, respectively. The total number of individuals is $N = n_{i'} + n_{j'}$. The assumptions of the **G** matrix is that the genotyped individuals represent the base population where the expected value of self-relatedness is 1, assuming no inbreeding, and that the mean relatedness of any one individual to the rest of the individuals is zero. Consequently, the expectation of diagonal elements of the **G** matrix is equal to the number of individuals assuming no inbreeding and the expectation of off-diagonal elements is $-1/(N-1)$ [e.g., 162]. Then the expectation of numerator in the r statistic is proportional to

$$E\left[\sum \mathrm{PEC}_{i'j'}\right] \propto -\frac{n_{i'}n_{j'}}{(N-1)},$$

and the expectation of denominator is the square root of the product between

$$\begin{aligned} E\left[\sum \mathrm{PEV}_{i'i'}\right] &\propto n_{i'} - 2 \cdot \frac{n_{i'}(n_{i'}-1)}{2} \cdot \frac{1}{(N-1)} \\ &= n_{i'} - \frac{n_{i'}^2 - n_{i'}}{N-1} \\ &= \frac{(N-1)n_{i'} - n_{i'}^2 + n_{i'}}{N-1} \\ &= \frac{Nn_{i'} - n_{i'} - n_{i'}^2 + n_{i'}}{N-1} \\ &= \frac{Nn_{i'} - n_{i'}^2}{N-1} \end{aligned}$$

and

$$\begin{aligned} E\left[\sum \mathrm{PEV}_{j'j'}\right] &\propto n_{j'} - 2 \cdot \frac{n_{j'}(n_{j'}-1)}{2} \cdot \frac{1}{(N-1)} \\ &= n_{j'} - \frac{n_{j'}^2 - n_{j'}}{N-1} \\ &= \frac{(N-1)n_{j'} - n_{j'}^2 + n_{j'}}{N-1} \\ &= \frac{Nn_{j'} - n_{j'} - n_{j'}^2 + n_{j'}}{N-1} \\ &= \frac{Nn_{j'} - n_{j'}^2}{N-1}, \end{aligned}$$

so that

$$\begin{aligned} E\left[\sum \mathrm{PEV}_{i'i'}\right] \cdot E\left[\sum \mathrm{PEV}_{j'j'}\right] &= \frac{Nn_{i'} - n_{i'}^2}{N-1} \cdot \frac{Nn_{j'} - n_{j'}^2}{N-1} \\ &= \frac{N^2 n_{i'}n_{j'} - Nn_{i'}n_{j'}^2 - Nn_{i'}^2 n_{j'} + n_{i'}^2 n_{j'}^2}{(N-1)^2}. \end{aligned}$$

Note that the first three terms in the numerator are equal to zero

$$\begin{aligned} N^2 n_{i'} n_{j'} - N n_{i'} n_{j'}^2 - N n_{i'}^2 n_{j'} &= N n_{i'} n_{j'} (N - n_{j'} - n_{i'}) \\ &= N n_{i'} n_{j'} \left[N - (n_{j'} + n_{i'}) \right] \\ &= N n_{i'} n_{j'} (N - N) \\ &= 0 \end{aligned}$$

because $N = n_{i'} + n_{j'}$. Therefore, the r statistic between units i' and j' is given by

$$\begin{aligned} \mathrm{r} &= \frac{E\left[\sum \mathrm{PEC}_{i'j'}\right]}{\sqrt{E\left[\sum \mathrm{PEV}_{i'i'}\right] \cdot E\left[\sum \mathrm{PEV}_{j'j'}\right]}} \\ &= \frac{-\frac{n_{i'} n_{j'}}{(N-1)}}{\sqrt{\frac{n_{i'}^2 n_{j'}^2}{(N-1)^2}}} \\ &= \frac{-\frac{n_{i'} n_{j'}}{(N-1)}}{\frac{n_{i'} n_{j'}}{(N-1)}} \\ &= -1. \end{aligned}$$

When $N = n_{i'} + n_{j'}$, this result holds regardless of relatedness level, connectedness level, and how individuals are partitioned into the two management units i' and j'. The partitioning of animals into two distinct units is particularly relevant in the context of genomic prediction where animals may be divided into training and testing sets. In this scenario, computing the connectedness between the two sets along the lines of Rincent et al. [57] and Isidro et al. [58] is potentially informative relative to expectations of the performance of resulting genomic predictors. The $\mathbf{G_s}$ or $\mathbf{G_{0.5}}$ matrix changes the expectation of off-diagonal elements to positive values and shifts the statistic by a constant as explained in the Methods section, yielding connectedness between units i' and j' of close to 1. Because scenarios 2 to 4 in the cattle dataset simulated two management units, the average of the r-statistic of pairs

of individuals in different management units was used to summarize connectedness in this study. Note that this is shown as lamb connectedness or individual connectedness in Kuehn et al. [14]. The two types of connectedness differ mainly by whether we take the average followed by the ratio (unit connectedness) or take the ratio first followed by the average (individual connectedness).

Appendix B

Changes of elements in PEV and PEC matrices

We further investigated how the $\mathbf{G}$ or $\mathbf{G_s}$ increased connectedness across management units using PEVD and r by examining the specific components in the scaled PEV matrix. For instance, disconnected "19F" and "13C" management units contained five full-sib individuals each. The following matrix contains the pedigree-based PEV for the 10 individuals.

$$\begin{bmatrix}
0.73 & 0.57 & 0.57 & 0.57 & 0.57 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.57 & 0.73 & 0.57 & 0.57 & 0.57 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.57 & 0.57 & 0.73 & 0.57 & 0.57 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.57 & 0.57 & 0.57 & 0.73 & 0.57 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.57 & 0.57 & 0.57 & 0.57 & 0.73 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.55 & 0.39 & 0.39 & 0.39 & 0.39 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.39 & 0.55 & 0.39 & 0.39 & 0.39 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.39 & 0.39 & 0.55 & 0.39 & 0.39 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.39 & 0.39 & 0.39 & 0.55 & 0.39 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.39 & 0.39 & 0.39 & 0.39 & 0.55
\end{bmatrix}$$
$$\underbrace{\qquad\qquad}_{19F} \qquad \underbrace{\qquad\qquad}_{13C}$$

The first five individuals belong to "19F" and the remaining individuals belong to "13C". Because there are no full-sib pairs across management units, off-diagonals are all zero. Thus,

PEVD and r of two individuals across management units are $0.73 + 0.55 - (2 \times 0) = 1.28$ and $0/(\sqrt{0.73} \times \sqrt{0.55}) = 0$, respectively.

When the **A** matrix is replaced with the **G** matrix, the PEV matrix becomes

$$\left[\begin{array}{cccccccccc}
0.30 & 0.19 & 0.19 & 0.19 & 0.19 & -0.01 & -0.01 & -0.01 & -0.01 & -0.01 \\
0.19 & 0.32 & 0.21 & 0.21 & 0.19 & -0.01 & -0.02 & -0.02 & -0.01 & -0.01 \\
0.19 & 0.21 & 0.32 & 0.19 & 0.20 & -0.01 & -0.01 & -0.01 & -0.01 & -0.01 \\
0.19 & 0.21 & 0.19 & 0.30 & 0.19 & -0.01 & -0.01 & -0.02 & -0.01 & -0.01 \\
0.19 & 0.19 & 0.20 & 0.19 & 0.30 & -0.01 & -0.01 & -0.02 & -0.01 & -0.01 \\
-0.01 & -0.01 & -0.01 & -0.01 & -0.01 & 0.24 & 0.14 & 0.14 & 0.15 & 0.14 \\
-0.01 & -0.02 & -0.01 & -0.01 & -0.01 & 0.14 & 0.24 & 0.16 & 0.14 & 0.13 \\
-0.01 & -0.02 & -0.01 & -0.02 & -0.02 & 0.14 & 0.16 & 0.26 & 0.14 & 0.14 \\
-0.01 & -0.01 & -0.01 & -0.01 & -0.01 & 0.15 & 0.14 & 0.14 & 0.25 & 0.14 \\
\underbrace{-0.01 \quad -0.01 \quad -0.01 \quad -0.01 \quad -0.01}_{19F} & & & & & \underbrace{0.14 \quad 0.13 \quad 0.14 \quad 0.14}_{13C} & & & & 0.24
\end{array}\right].$$

Average genomic relationships within management units were 0.419 and 0.440 for individuals in "19F" and "13C", respectively, whereas across management unit genomic relationships were -0.09. The off-diagonals of zeros in pedigree-based PEV were replaced with small negative values. Although the diagonal elements within management units are not all equal because of Mendelian sampling, PEVD between the first individuals from respective management units are $0.30 + 0.24 - (2 \times -0.01) = 0.56$. Given that off-diagonal elements are negligible, the rate of PEVD reduction from shifting from **A** to **G** is almost 50% with the most of difference coming from decreased PEV in the diagonals. Specifically, the rates of PEV reduction (diagonals) from **A** to **G** were 59% and 56% for the first individuals in "19F" and "13C", respectively. The rates of PEC reduction cannot be defined since all of the off-diagonal elements are zeros in **A**. Note that the r statistic using the **G** matrix

does not yield increased estimates of connectedness compared with that using **A** because $-0.01/(\sqrt{(}0.30) \times \sqrt{0.24}) = -0.04$. Now consider the $\mathbf{G_s}$ matrix

$$\underbrace{\begin{bmatrix} 0.68 & 0.56 & 0.56 & 0.55 & 0.56 \\ 0.56 & 0.69 & 0.58 & 0.57 & 0.56 \\ 0.56 & 0.58 & 0.69 & 0.56 & 0.56 \\ 0.55 & 0.57 & 0.56 & 0.67 & 0.56 \\ 0.56 & 0.56 & 0.56 & 0.56 & 0.67 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0.32 \\ 0.32 & 0.31 & 0.32 & 0.32 & 0.32 \\ 0.32 & 0.31 & 0.32 & 0.31 & 0.31 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0.32 \end{bmatrix}}_{19F} \underbrace{\begin{bmatrix} 0.32 & 0.32 & 0.32 & 0.32 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.31 & 0.31 \\ 0.61 & 0.50 & 0.50 & 0.51 & 0.49 \\ 0.50 & 0.61 & 0.51 & 0.49 & 0.48 \\ 0.50 & 0.51 & 0.63 & 0.50 & 0.50 \\ 0.51 & 0.49 & 0.50 & 0.62 & 0.49 \\ 0.49 & 0.48 & 0.50 & 0.49 & 0.60 \end{bmatrix}}_{13C}.$$

Here the PEV matrix within management units more closely resembles those from pedigree-based PEV. In addition, the negative elements across management units PEV were replaced with positive values. With use of $\mathbf{G_s}$, PEVD and r between the first individuals from the two management units are $0.68 + 0.61 - (2 \times 0.32) = 0.65$ and $0.32/(\sqrt{0.68} \times \sqrt{0.61}) = 0.50$, respectively. When the scaled genomic relationship matrix $\mathbf{G_s}$ is used, r yields an increased connectedness estimate as compared to using pedigree-based relationships.

Likewise, a subsequent example with "19F" and "36F" which can be viewed as connected management units (Figure 3.1). Each management unit contains five full-sibs as in the

previous case. The following matrix of pedigree-based PEV includes these 10 individuals.

$$\begin{bmatrix} 0.73 & 0.57 & 0.57 & 0.57 & 0.57 & 0.50 & 0.50 & 0.50 & 0.50 & 0.50 \\ 0.57 & 0.73 & 0.57 & 0.57 & 0.57 & 0.50 & 0.50 & 0.50 & 0.50 & 0.50 \\ 0.57 & 0.57 & 0.73 & 0.57 & 0.57 & 0.50 & 0.50 & 0.50 & 0.50 & 0.50 \\ 0.57 & 0.57 & 0.57 & 0.73 & 0.57 & 0.50 & 0.50 & 0.50 & 0.50 & 0.50 \\ 0.57 & 0.57 & 0.57 & 0.57 & 0.73 & 0.50 & 0.50 & 0.50 & 0.50 & 0.50 \\ 0.50 & 0.50 & 0.50 & 0.50 & 0.50 & 0.73 & 0.57 & 0.57 & 0.57 & 0.57 \\ 0.50 & 0.50 & 0.50 & 0.50 & 0.50 & 0.57 & 0.73 & 0.57 & 0.57 & 0.57 \\ 0.50 & 0.50 & 0.50 & 0.50 & 0.50 & 0.57 & 0.57 & 0.73 & 0.57 & 0.57 \\ 0.50 & 0.50 & 0.50 & 0.50 & 0.50 & 0.57 & 0.57 & 0.57 & 0.73 & 0.57 \\ 0.50 & 0.50 & 0.50 & 0.50 & 0.50 & 0.57 & 0.57 & 0.57 & 0.57 & 0.73 \end{bmatrix}$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad}_{19F} \quad \underbrace{\qquad\qquad\qquad\qquad\qquad}_{36F}$$

The first five individuals belong to "19F" and the remaining individuals belong to "36F". In this case, off-diagonals are non-zero due to the presence of shared full-sib information across management units. Here PEVD and r of two individuals is $0.73 + 0.73 - (2 \times 0.50) = 0.46$ and $0.50/(\sqrt{0.73} \times \sqrt{0.73}) = 0.68$, respectively. Relative to the pedigree-based non-full-sib comparison, a significant increase in connectedness was observed. The majority of the increase in connectedness is due to increased PEC between individuals.

The following is the PEV matrix when pedigree is substituted with the genome-wide markers

from **G**.

$$\begin{bmatrix} 0.30 & 0.19 & 0.19 & 0.19 & 0.19 & 0.19 & 0.18 & 0.18 & 0.18 & 0.19 \\ 0.19 & 0.32 & 0.21 & 0.21 & 0.19 & 0.18 & 0.18 & 0.18 & 0.19 & 0.19 \\ 0.19 & 0.21 & 0.32 & 0.19 & 0.20 & 0.18 & 0.18 & 0.18 & 0.20 & 0.19 \\ 0.19 & 0.21 & 0.19 & 0.30 & 0.19 & 0.18 & 0.17 & 0.19 & 0.18 & 0.18 \\ 0.19 & 0.19 & 0.20 & 0.19 & 0.30 & 0.18 & 0.17 & 0.18 & 0.18 & 0.19 \\ 0.19 & 0.18 & 0.18 & 0.18 & 0.18 & 0.31 & 0.20 & 0.19 & 0.21 & 0.20 \\ 0.18 & 0.18 & 0.18 & 0.17 & 0.17 & 0.20 & 0.31 & 0.18 & 0.21 & 0.20 \\ 0.18 & 0.18 & 0.18 & 0.19 & 0.18 & 0.19 & 0.18 & 0.29 & 0.20 & 0.20 \\ 0.18 & 0.19 & 0.20 & 0.18 & 0.18 & 0.21 & 0.21 & 0.20 & 0.32 & 0.21 \\ 0.19 & 0.19 & 0.19 & 0.18 & 0.19 & 0.20 & 0.20 & 0.20 & 0.21 & 0.31 \end{bmatrix}$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad}_{19F} \quad \underbrace{\qquad\qquad\qquad\qquad\qquad}_{36F}$$

Average genomic relationship within "36F" was 0.465 whereas average genomic relationship across "19F" was 0.419. The off-diagonals no longer have small negative values and all elements of PEV were reduced in comparison to those using **A**. Here PEVD and r between the first individuals from the two management units are $0.30 + 0.31 - (2 \times 0.19) = 0.23$ and $0.19/(\sqrt{0.30} \times \sqrt{0.31}) = 0.62$, respectively. Again, while genomic information increased estimates of connectedness as measured by PEVD, that was not the case for r. The reduction in PEVD from **A** to **G** was about 50% and both diagonals and off-diagonals contributed to increasing connectedness estimates. In particular, the rates of PEV reduction (diagonals) from **A** to **G** were 59% and 58% for the first individuals in "19F" and "36F", respectively. The reduction in PEC due to the use of **G** was 62% for these two individuals, which was larger than the reduction of the diagonals. This contributed to the unexpected results for r because this statistic is based on a ratio. Now consider use of the scaled genomic relationship

matrix, $\mathbf{G_s}$, which yielded the following PEV matrix

$$\left[\begin{array}{cccccccccc}
0.68 & 0.56 & 0.56 & 0.55 & 0.56 & 0.55 & 0.55 & 0.54 & 0.54 & 0.56 \\
0.56 & 0.69 & 0.58 & 0.57 & 0.56 & 0.54 & 0.54 & 0.54 & 0.55 & 0.55 \\
0.56 & 0.58 & 0.69 & 0.56 & 0.56 & 0.54 & 0.55 & 0.54 & 0.56 & 0.55 \\
0.55 & 0.57 & 0.56 & 0.67 & 0.56 & 0.54 & 0.53 & 0.55 & 0.54 & 0.55 \\
0.56 & 0.56 & 0.56 & 0.56 & 0.67 & 0.55 & 0.53 & 0.54 & 0.54 & 0.55 \\
0.55 & 0.54 & 0.54 & 0.54 & 0.55 & 0.69 & 0.57 & 0.55 & 0.58 & 0.56 \\
0.55 & 0.54 & 0.55 & 0.53 & 0.53 & 0.57 & 0.68 & 0.55 & 0.57 & 0.56 \\
0.54 & 0.54 & 0.54 & 0.55 & 0.54 & 0.55 & 0.55 & 0.67 & 0.57 & 0.57 \\
0.54 & 0.55 & 0.56 & 0.54 & 0.54 & 0.58 & 0.57 & 0.57 & 0.70 & 0.58 \\
\underbrace{0.56}_{19F} & 0.55 & 0.55 & 0.55 & 0.55 & \underbrace{0.56}_{36F} & 0.56 & 0.57 & 0.58 & 0.69
\end{array}\right].$$

The PEV matrix including both within and across management units are more analogous to those of the pedigree-based PEV. Here PEVD and r between the first individuals from the two management units are $0.68 + 0.69 - (2 \times 0.55) = 0.27$ and $0.55/(\sqrt{0.68} \times \sqrt{0.69}) = 0.80$, respectively. The result reaffirms that scaling $\mathbf{G}$ to be on the same scale as $\mathbf{A}$, genomic relatedness leads to an increases in the ratio sensitive connectedness statistic, a change consistent with the other statistics. Collectively, these particular examples, suggest that genomic information provided by $\mathbf{G}$ or $\mathbf{G_s}$ changes the estimates of relationship coefficients among animals and refines the estimates of connectedness.

Appendix C

Description of phenotypes

Table C.1: Description of phenotypes used.

Phenotype Label	Category	Description	Publication
Flowering time at Arkansas (Fla)	Flowering time	Number of days until the inflorescence is 50% emerged from the flag leaf counted from the day of planting. Measured in Arkansas, USA.	Zhao et al. [122]
Flowering time at Faridpur (Flf)	Flowering time	Number of days until the inflorescence is 50% emerged from the flag leaf counted from the day of transplanting. Measured in Faridpur, BD.	Zhao et al. [122]
Flowering time at Aberdeen (Flb)	Flowering time	Number of days until the inflorescence is 50% emerged from the flag leaf counted from the day of planting. Measured in Aberdeen, UK.	Zhao et al. [122]
FT ratio of Arkansas/Aberdeen (Flra)	Flowering time	Ratio of days to heading in Arkansas to days to heading in Aberdeen. Metric used to describe photoperiod sensitivity.	Zhao et al. [122]
FT ratio of Faridpur/Aberdeen (Flfa)	Flowering time	Ratio of days to heading in Faridpur to days to heading in Aberdeen. Metric used to describe photoperiod sensitivity.	Zhao et al. [122]
Year07 Flowering time at Arkansas (Fla7)	Flowering time	Number of days until the inflorescence is 50% emerged from the flag leaf counted from the day of planting. Measured in Arkansas in 2007.	Zhao et al. [122]
Year06 Flowering time at Arkansas (Fla6)	Flowering time	Number of days until the inflorescence is 50% emerged from the flag leaf counted from the day of planting. Measured in Arkansas in 2006.	Zhao et al. [122]
Seed length (Sl)	Grain morphology	Length of the seed with hull (palea and lemma)	Zhao et al. [122]
Seed width (Sw)	Grain morphology	Width of the seed with hull (palea and lemma)	Zhao et al. [122]
Seed volume (Sv)	Grain morphology	Volume of the seed with hull (palea and lemma)	Zhao et al. [122]
Seed surface area (Ssa)	Grain morphology	Surface area of the seed with hull (palea and lemma)	Zhao et al. [122]
Brown rice seed length (Bsl)	Grain morphology	Length of the unpolished rice grain (dehulled seed)	Zhao et al. [122]
Brown rice seed width (Bsw)	Grain morphology	Width of the unpolished rice grain (dehulled seed)	Zhao et al. [122]
Brown rice surface area (Bsa)	Grain morphology	Surface area of the unpolished rice grain (dehulled seed)	Zhao et al. [122]
Brown rice volume (Bsl)	Grain morphology	Volume of the unpolished rice grain (dehulled seed)	Zhao et al. [122]
Seed length/width ratio (Slwr)	Grain morphology	Ratio of seed length/ seed width (with hull)	Zhao et al. [122]
Brown rice length/width ratio (Blwr)	Grain morphology	Ratio of unpolished rice grain length/grain width (dehulled seed)	Zhao et al. [122]
Grain length McCouch2016 (Glmc)	Grain morphology	Length of the seed with hull (palea and lemma). Reported by McCouch et al (JSBD)	McCouch et al. [123]
Culm habit (Cuh)	Morphology	Average culm angle of plants at maturity	Zhao et al. [122]
Flag leaf length (Fll)	Morphology	Length of the flag leaf measured from leaf base to leaf tip (cm)	Zhao et al. [122]
Flag leaf width (Flw)	Morphology	Width of the flag leaf measured at the widest portion of flag leaf lamina (cm)	Zhao et al. [122]
Plant height (Plh)	Morphology	Height of plant from soil surface to tip of main panicle (inflorescence) (cm)	Zhao et al. [122]
Shoot BM Control (Shc)	Morphology	Shoot dry weight (g) for 28 day old plants in control.	Campbell et al. [124]
Shoot BM Salt (Shs)	Morphology	Shoot dry weight (g) for 28 day old plants 9 dS salt.	Campbell et al. [124]
Root BM Control (Rbc)	Morphology	Root dry weight (g) for 28 day old plants in control.	Campbell et al. [124]
Root BM Salt (Rbsl)	Morphology	Root dry weight (g) for 28 day old plants in 9 dS salt.	Campbell et al. [124]
Tiller No Salt (Tbs)	Morphology	Tiller number for 28 day old plants in 9 dS salt.	Unpublished. Study described in Campbell et al. [124]
Tiller No Control (Tbc)	Morphology	Tiller number for 28 day old plants in control.	Unpublished. Study described in Campbell et al. [124]
Ht. Lig Salt (Hls)	Morphology	Height (cm) from the soil to the ligule of the newest expanded leaf in salt. Measured on 28 day old plants.	Unpublished. Study described in Campbell et al. [124]
Ht. Lig Control (Hlc)	Morphology	Height (cm) from the soil to the ligule of the newest expanded leaf in salt. Measured on 28 day old plants.	Unpublished. Study described in Campbell et al. [124]
Ht FE Salt (Hfs)	Morphology	Height (cm) from the soil to the tip of the newest expanded leaf in salt.	Unpublished. Study described in Campbell et al. [124]
Ht FE Control (Hfc)	Morphology	Height (cm) from the soil to the tip of the newest expanded leaf in control. Measured on 28 day old plants.	Unpublished. Study described in Campbell et al. [124]
Na K Shoot (Ks)	Ionic components of salt stress	Shoot sodium to potassium ratio (10') in salt. Measured on 28 day old plants after 14 days of 9dS salt stress.	Campbell et al. [124]
Na Shoot (Nas)	Ionic components of salt stress	Shoot sodium content (mmol/g dry wt.) in salt. Measured on 28 day old plants after 14 days of 9dS salt stress.	Campbell et al. [124]
K Shoot Salt (Kss)	Ionic components of salt stress	Shoot potassium content (mmol/g dry wt.) in salt. Measured on 28 day old plants after 14 days of 9dS salt stress.	Campbell et al. [124]
Na K Root (Kr)	Ionic components of salt stress	Root sodium to potassium ratio (10') in salt. Measured on 28 day old plants after 14 days of 9dS salt stress.	Campbell et al. [124]
Na Root (Nad)	Ionic components of salt stress	Root sodium content (mmol/g dry wt.) in salt. Measured on 28 day old plants after 14 days of 9dS salt stress.	Campbell et al. [124]
K Root Salt (Krsl)	Ionic components of salt stress	Root potassium content (mmol/g dry wt.) in salt. Measured on 28 day old plants after 14 days of 9dS salt stress.	Campbell et al. [124]
Shoot BM Ratio (Slsr)	Morphological salt response	Ratio of the LSMeans for shoot dry in salt over control. Measured on 28 day old plants after 14 days of 9dS salt stress.	Campbell et al. [124]
Root BM Ratio (Rbr)	Morphological salt response	Ratio of the LSMeans for root dry in salt over control. Measured on 28 day old plants after 14 days of 9dS salt stress.	Campbell et al. [124]
Tiller No Ratio (Tbr)	Morphological salt response	Ratio of the tiller number in salt over tiller number in control. Measured on 28 day old plants after 14 days of 9dS salt stress.	Unpublished. Study described in Campbell et al. [124]
Ht Lig Ratio (Hlr)	Morphological salt response	Ratio of height from the soil to the ligule of the newest expanded leaf in salt over control. Measured on 28 day old plants after 14 days of 9dS salt stress.	Unpublished. Study described in Campbell et al. [124]
Ht FE Ratio (Hfr)	Morphological salt response	Ratio of height (cm) from the soil to the ligule of the newest expanded leaf in salt over control. Measured on 28 day old plants after 14 days of 9dS salt stress.	Unpublished. Study described in Campbell et al. [124]
Panicle number per plant (Pnu)	Yield	Average number of panicles (inflorescences) per plant	Zhao et al. [122]
Panicle length (Pal)	Yield	Length of panicle (inflorescence) from the base to the tip (cm)	Zhao et al. [122]
Primary panicle branch number (Ppn)	Yield	Number of primary branches along the panicle (inflorescence)	Zhao et al. [122]
Seed number per panicle (Sspp)	Yield	Number of seeds per panicle (inflorescence), determined by counting the number of filled spikelets along the main panicle	Zhao et al. [122]
Panicle fertility (Paf)	Yield	Percent of spikelets that filled and produced seeds determined as the ratio of seeds per panicle/spikelets per panicle	Zhao et al. [122]