

RESEARCH ARTICLE

# SNP variable selection by generalized graph domination

Shuzhen Sun<sup>1,2</sup>, Zhuqi Miao<sup>3</sup>, Blaise Ratcliffe<sup>2</sup>, Polly Campbell<sup>4,5</sup>, Bret Pasch<sup>6</sup>, Yousry A. El-Kassaby<sup>2</sup>, Balabhaskar Balasundaram<sup>7</sup>, Charles Chen<sup>1\*</sup>

**1** Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, United States of America, **2** Department of Forest and Conservation Sciences, Faculty of Forestry, The University of British Columbia, Vancouver, B.C. Canada, **3** Center for Health Systems Innovation, Oklahoma State University, Stillwater, United States of America, **4** Department of Integrative Biology, Oklahoma State University, Stillwater, United States of America, **5** Department of Evolution, Ecology and Organismal Biology, University of California, Riverside, Riverside, United States of America, **6** Department of Biological Sciences, Northern Arizona University, Flagstaff, United States of America, **7** School of Industrial Engineering and Management, Oklahoma State University, Stillwater, United States of America

\* [charles.chen@okstate.edu](mailto:charles.chen@okstate.edu)



## Abstract

### OPEN ACCESS

**Citation:** Sun S, Miao Z, Ratcliffe B, Campbell P, Pasch B, El-Kassaby YA, et al. (2019) SNP variable selection by generalized graph domination. PLoS ONE 14(1): e0203242. <https://doi.org/10.1371/journal.pone.0203242>

**Editor:** Siamak Yassemi, University of Tehran, ISLAMIC REPUBLIC OF IRAN

**Received:** August 15, 2018

**Accepted:** January 8, 2019

**Published:** January 24, 2019

**Copyright:** © 2019 Sun et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** SNP-SELECT and the SNP datasets used in this manuscript are available on the GitHub repository (<https://github.com/transgenomicsosu/SNP-SELECT>). For additional details regarding the original source code developed for SNP-SELECT, as well as the description of the datasets, see the README.md.

**Funding:** This research is funded by Oklahoma Wheat Research Foundation, OCAST (PS15-011) and NSF-MRI 1626257 (CC), NSF-IOS 1558109 (CC and PC), NSF-CMMI 1404971 (BB), and a fellowship from the Cornell Lab of Ornithology

## Background

High-throughput sequencing technology has revolutionized both medical and biological research by generating exceedingly large numbers of genetic variants. The resulting datasets share a number of common characteristics that might lead to poor generalization capacity. Concerns include noise accumulated due to the large number of predictors, sparse information regarding the  $p \gg n$  problem, and overfitting and model mis-identification resulting from spurious collinearity. Additionally, complex correlation patterns are present among variables. As a consequence, reliable variable selection techniques play a pivotal role in predictive analysis, generalization capability, and robustness in clustering, as well as interpretability of the derived models.

## Methods and findings

$K$ -dominating set, a parameterized graph-theoretic generalization model, was used to model SNP (single nucleotide polymorphism) data as a similarity network and searched for representative SNP variables. In particular, each SNP was represented as a vertex in the graph, (dis)similarity measures such as correlation coefficients or pairwise linkage disequilibrium were estimated to describe the relationship between each pair of SNPs; a pair of vertices are adjacent, i.e. joined by an edge, if the pairwise similarity measure exceeds a user-specified threshold. A minimum  $k$ -dominating set in the SNP graph was then made as the smallest subset such that every SNP that is excluded from the subset has at least  $k$  neighbors in the selected ones. The strength of  $k$ -dominating set selection in identifying independent variables, and in culling representative variables that are highly correlated with others, was demonstrated by a simulated dataset. The advantages of  $k$ -dominating set variable selection were also illustrated in two applications: pedigree reconstruction using SNP profiles of 1,372 Douglas-fir trees, and species delineation for 226 grasshopper mouse

(BP). The work presented in this report also reflects the support from the USDA HATCH project OKL03011 (CC). SS, BR, YAE and CC acknowledge cash funding for this research from Genome Canada, Genome Alberta through Alberta Economic Trade and Development, Genome British Columbia, the University of Alberta and University of Calgary and others, including the Alberta forest industry in support of the Resilient Forests (RESFOR): Climate, Pests & Policy- Genomic Applications project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

samples. A C++ source code that implements SNP-SELECT and uses Gurobi optimization solver for the  $k$ -dominating set variable selection is available (<https://github.com/transgenomicsosu/SNP-SELECT>).

## Introduction

With the rapid advancement of DNA sequencing technology, the volume and dimension of biological and medical data have been increasing at an unprecedented rate. Accompanying such high volume genetic data, the ‘curse of dimensionality’ has challenged the validity of statistical methods that do not scale to massive data. Statistical accuracy, model interpretability and computational efficiency could be significantly impacted, especially when the number of predictors is much greater than sample size [1]. For instance in high dimensional classification, conventional classification rules using all variables perform no better than random guess for small sample sizes [2]; and in omics data analysis where the ultimate goal is to identify a small number of predictors (biomarkers, metabolites or genes), the correlation structure among predictors in the biology of the experiment often complicates biomarker identification [3]. Sources for these unsatisfied algorithmic performance could be the result of model noise accumulation in high dimension, incidental correlation between residual errors and some predictors, and the spurious collinearity that causes over-fitting and mis-identification of models [4–6], making variable selection a practical solution for “large  $p$  small  $n$ ” data [7, 8].

Magnitude and significance of linkage disequilibrium (LD) in the genome markedly varies between populations [9, 10], causing unexpected multi-collinearity that leads to unstable estimates of genetic parameters [11]. By reducing correlation in SNP predictors, Song et al. [12] and others showed that with a selected subset, comparable predictability for complex traits like grain yield and milk yield could be achieved [11, 13–16]. Results from Weigel et al. (2009) [17] further suggest that, not only compatible prediction accuracy could be derived from a much smaller, evenly spaced SNP subset, but the standard deviation of prediction accuracy reduced. Crucial to both analyzing and interpreting high dimensional SNP datasets, significant effort has been directed towards exploring variable selection processes by removing features that might be either redundant or irrelevant to the problem, for better predictability, or computational efficiency and informativeness [18]. This effort includes the logistic regression method [19], the penalized regression method [1, 20, 21], partial least squares regression (PLSR) [22], sure independence screening strategy [23], multi-stage regression methods [24], sorted l-one penalized estimation (SLOPE) via convex optimization [25], recurrent relative variable importance measure (r2VIM) [26], to name a few. However, these methods were designed to reduce variables from a statistical perspective in order to ease the process of prediction or assist GWAS (genome-wide association study) analysis, in which knowledge of phenotypic data is required.

In the era of population genomics [27], many *Fst*-based genome-scan methods utilize large datasets such as SNP chips or genome complexity reduction approaches like RAD tags [28] and genotyping-by-sequencing (GBS) [29, 30], to estimate genetic parameters [31]. Identifying adaptive evolution and candidate genomic regions under selection is increasingly feasible, thanks to the development of sophisticated analytical tools for genome-scale polymorphism data [32–35]. Given the data volume, most of these Bayesian approaches suffer from extended computational time requirement [31] due to tedious numerical approximation procedures like Markov chain Monte Carlo (MCMC) [34] or reverse jump (RJ)-MCMC [33]. Furthermore,

accurate inference of demographic parameters such as effective population sizes, migration rates, and divergence times between populations depends largely on the use of neutral marker data [36–38]. In other words, SNP variable selection methods without the use of phenotypic data are desirable for the purpose of reducing the bias caused by confounding variables, for minimizing computational load, and for avoiding the potential problem of allele frequency correlations in, for example, the Lewontin and Krakauer (LK) test [31, 39].

In this paper, we present SNP-SELECT, a variable selection algorithm based on a graph-theory approach that uses generalized dominating sets for a large volume of SNP data without the use of phenotypes. Application of graph theory to variable selection or data reduction has been seen in many data mining applications [40–42]. Typically, this involves clustering the data points into groups and using one point to represent each cluster, from which the network clustering [43] procedure would derive a much smaller number of clusters, resulting in variable selection. In our cases, data points (SNPs) are represented by vertices and an edge exists if two data points (two SNPs) are similar or related in a certain way (i.e., in LD or in correlation). We show the use of LD with an example; it is, however, important to note that the similarity criterion used to construct the network model can be based on any relationship measurement. The advantage and robustness of SNP-SELECT is also demonstrated with simulated datasets, and with empirical datasets for a Douglas-fir (*Pseudotsuga menziesii*) breeding population and for populations of three grasshopper mouse species (*Onychomys* spp.).

## Material and methods

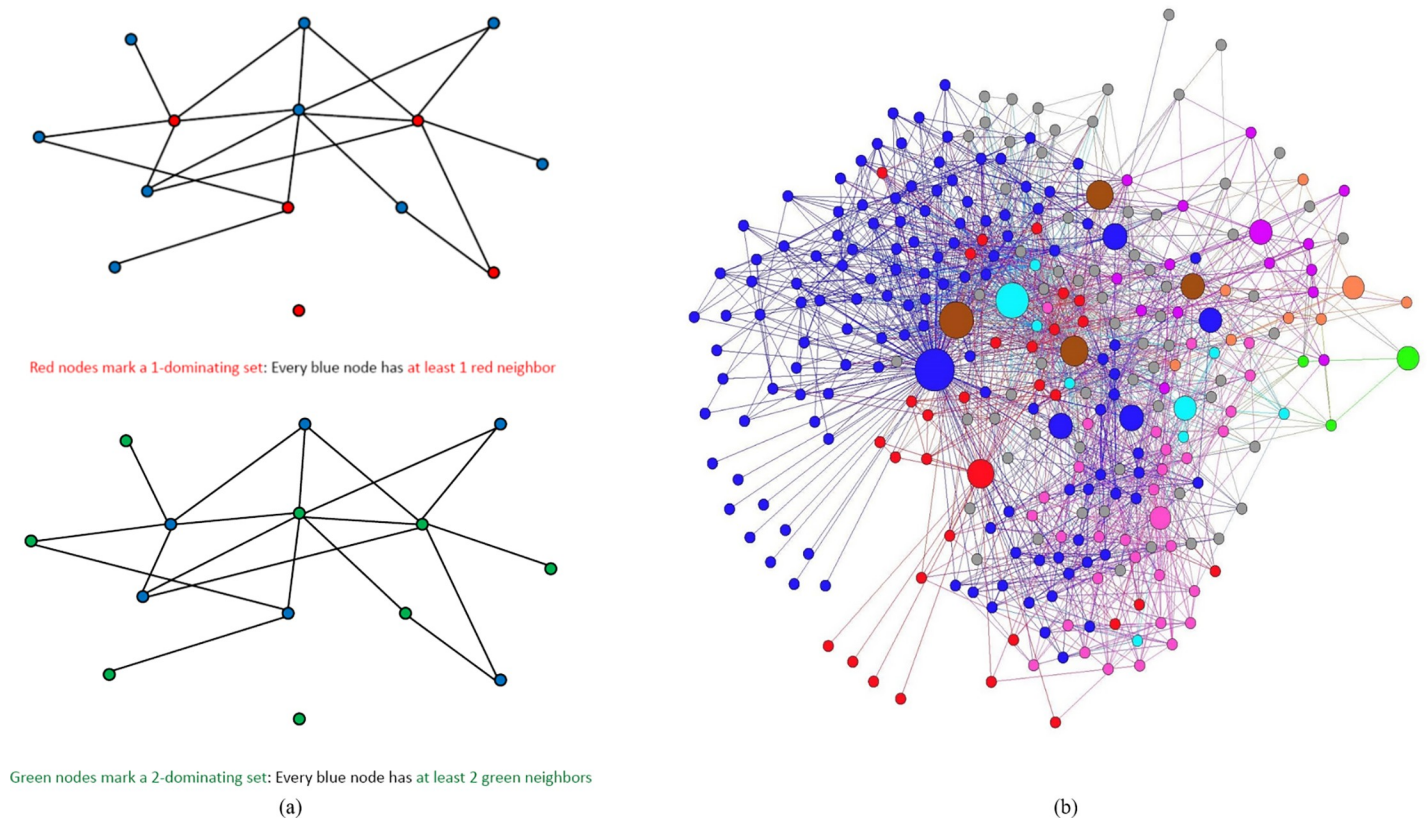
### Generalized graph domination

Let  $G = (V, E)$  be a graph with vertex set  $V$  and edge set  $E \subseteq [V]^2$  (see [44] for basic graph theory concepts and notations). The *open neighborhood* of a vertex  $v$  is the set  $N(v)$  of vertices adjacent to vertex  $v$ . Note that  $v \notin N(v)$  and the *closed neighborhood* of vertex  $v$  is denoted by  $N[v] = \{v\} \cup N(v)$ .

**Definition 1** [45] Given a positive integer  $k$  and a graph  $G = (V, E)$ , a subset of vertices  $D$  is said to be  $k$ -dominating if  $|N(v) \cap D| \geq k$  for every vertex  $v \notin D$ .

If  $D$  is a  $k$ -dominating set, then every vertex in  $V - D$  is said to be  $k$ -dominated. A *minimum  $k$ -dominating set* is one of smallest cardinality in the graph and this cardinality is called the  *$k$ -domination number* of the graph, denoted as  $\gamma_k(G)$ . Note that the  $k$ -domination number of a graph increases as parameter  $k$  increases and the model becomes more restrictive as more neighbors are needed for each vertex outside the set to be  $k$ -dominated. Hence, every 2-dominating set is also a 1-dominating set, but the converse is not true. Intuitively, as the parameter  $k$  increases, we expect the  $k$ -dominating set to be a more reliable representation of the dataset as each point has at least  $k$  similar points in the  $k$ -dominating set. Hence, the choice of  $k$  must balance two conflicting criteria: solution fidelity (how well the dataset is represented) and solution size (how many data points are selected). To illustrate, graphic presentations of  $k$ -dominating sets for  $k = 1$  and  $k = 2$  were showed in Fig 1(A); and Fig 1(B) illuminated 1-dominating set using neural network data for the nematode, *C. elegans* [46, 47]. Neurons are represented by vertices in this neural network and as long as two neurons communicate with each other, an edge exists between them. The big dots in Fig 1(B) mark a 1-dominating set, and all the small dots (vertices) have at least one neighbor of the same color, which identifies the cluster.

Clustering a graph via  $k$ -dominating sets, especially with  $k = 1$ , is a popular technique in telecommunication and wireless networks [48]. If  $D$  is a 1-dominating set, then for each vertex  $v \in D$  the closed neighborhood  $N[v]$  forms a cluster that altogether cover  $V$ . Since by definition, every vertex not in the 1-dominating set has a neighbor in it and hence, is assigned to at least



**Fig 1.** (a) Illustration of 1-dominating set and 2-dominating set; (b) Illustration of 1-dominating set using the neural network data of *C. elegans* [46, 47]: the big nodes mark a 1-dominating set, and all the small nodes have at least 1 same color neighbor.

<https://doi.org/10.1371/journal.pone.0203242.g001>

one cluster. Since the problem of finding a minimum  $k$ -dominating set is NP-hard [49], heuristic approaches and approximation algorithms have been proposed to find a small  $k$ -dominating set in the given graph [50]. However, the approach employed in this article to solve this combinatorial optimization problem was to formulate it as an integer program [51], implement and solve it using a state-of-the-art solver that employs a branch-and-cut algorithm with built-in primal heuristics and other presolve reductions among others. Given a positive integer  $k$  and a graph  $G = (V, E)$ , the problem of finding a minimum  $k$ -dominating set can be formulated as the following linear integer program in binary variables.

$$\gamma_k(G) = \min \sum_{i \in V} x_i$$

subject to:

$$\sum_{j \in N(i)} x_j \geq k(1 - x_i), \quad \forall i \in V$$

$$x_i \in \{0, 1\}, \quad \forall i \in V$$

In any feasible solution  $x$  to this formulation, the binary variable  $x_i = 1$  if and only if vertex  $i$  is included in the  $k$ -dominating set  $D$ , which is given by  $D = \{i \in V: x_i = 1\}$ . The constraints ensure that if a vertex  $i$  is excluded from the  $k$ -dominating set  $D$ , i.e.  $x_i = 0$ , at least  $k$  of its neighbors must be included.

## Pairwise relationship between SNPs

The pairwise relationship (similarity or distance) between SNP variables primarily determines the structure of the graph  $G$ , and different ways for quantifying the pairwise relationship can influence the structure of the graph  $G$ , especially the sets of edges. Currently, many methods exist to measure the pairwise relationship of SNPs, for example, Hamming distance [52], mutual information [53, 54], allele sharing index [55, 56], and linkage disequilibrium (LD) [57–59], to name a few. We chose to use the frequently used LD approach to describe the pairwise relationship between SNP variables in this study, although the proposed approach continues to work with other similarity measures as well. The square of correlation coefficients ( $r^2$ ) for SNP variables were calculated to represent the values in the LD matrix (refer to [60] for the details). Since the gametic phase of haplotype frequencies for each pair of SNPs are unknown, the expectation maximization algorithm [61] was applied to infer the haplotype frequencies in the LD calculation.

With a user-defined threshold ( $\lambda$ ), an edge exists only if the pairwise relationship between the two SNPs (vertices) is greater than  $\lambda$ . Thus, for any given pairwise relationship measurement, as  $\lambda$  increases, the number of edges in the graph decreases, and consequently the number of isolated (independent) vertices in a graph can increase. For any positive integer  $k$ , an isolated vertex in the graph cannot be  $k$ -dominated by any other vertex, and must be included in any  $k$ -dominating set. In fact, this observation holds more generally for any vertex with fewer than  $k$  neighbors in the graph.

## Scheme of SNP-SELECT

The details of SNP-SELECT are summarized as follows:

**Step 1:** Construct a graph model  $G = (V, E)$ : Let  $V$  be the set of all SNPs and  $E$  is initially empty;

**Step 2:** Calculate linkage disequilibrium  $w_{ij}$  for each pair of SNPs  $i, j \in V$ ;

**Step 3:** An edge between SNPs  $i$  and  $j$  is created if  $w_{ij} > \lambda$ ;

**Step 4:** Identify isolated SNPs  $I \leftarrow \{i \in V: N(i) = \emptyset\}$ ;

**Step 5:** Find a minimum  $k$ -dominating set in  $G - I$ .

All experiments/analyses reported in this article were conducted on a 64-bit Linux compute node of a high performance computing cluster with 96GB RAM and Intel Xeon E5620 2.40GHz processor. The algorithm was implemented using C++, and the integer programming formulation for the minimum  $k$ -dominating set problem was solved using the Gurobi<sup>TM</sup> optimizer 6.0 with default settings [62]. Given a running time limit, Gurobi<sup>TM</sup> either returned an optimal solution, or a feasible solution with a gap to a lower bound on the optimal solution. Experiments/analyses reported in this study were performed with a 1-hour running time limit for Gurobi<sup>TM</sup>. The solution returned by Gurobi<sup>TM</sup> was used to identify the representative subset of the original dataset.

In our preliminary analyses, we found that when  $\lambda$  is small, e.g.  $\lambda < 0.2$ , the graph model tends to be very dense with an extremely large number of edges. When several thousands of SNPs are involved, such graphs can exceed memory limits during computation and result in a memory crash, before a feasible solution can be derived. Also, very small thresholds may not necessarily be realistic to capture similarities between SNPs. To address this issue, a stepwise search was implemented in SNP-SELECT for large SNP datasets as follows:

**Step 1:** Construct a threshold set  $T = \{\lambda_1, \lambda_2, \dots, \lambda_L\}$ , where  $\lambda_1 > \lambda_2 > \dots > \lambda_L$ , and  $\lambda_L$  is the desired threshold,  $\lambda_L \leftarrow \lambda$ , and  $\lambda_h - \lambda_{h+1}$  equals a predefined step; Let  $h = 1$ , and  $V_1$  be the set of all SNPs;

**Step 2:** Construct  $G_h$  on  $V_h$  using  $\lambda_h$ ;

**Step 3:** Identify isolated SNPs  $I_h \leftarrow \{i \in V_h: N(i) = \emptyset\}$ ;

**Step 4:** Find a minimum (or a small)  $k$ -dominating set  $S_h$  in  $G_h - I_h$ , let  $Y_h \leftarrow S_h \cup I_h$ ;

**Step 5:** If  $h = L$ , return  $Y_h$ , STOP; else  $V_{h+1} \leftarrow Y_h$ ,  $h \leftarrow h+1$ , go to **Step 2**.

In brief, this step-wise search of SNP-SELECT first finds a minimum  $k$ -dominating set  $Y_1$  (or the best solution available) on a graph model based on a larger threshold. Then the threshold is lowered to focus on the graph induced by  $Y_1$ . The data size of current step is the output of previous step. This process is repeated until a desired low threshold is reached. The feature selection problem of large datasets is thus solved by iteratively reducing the value of threshold.

### Simulation studies

To demonstrate the capacity of the  $k$ -dominating set algorithm to identify independent variables, and to select proxy variables among highly correlated ones, a simulated dataset that included 10 synthetic undirected networks with  $n = 1000$  vertices were used to represent SNPs. In this synthetic network dataset, the pairwise relationships between SNPs (vertices), the weighted edge ( $w_{ij}$ ) between each pair of vertices ( $i, j$ ), were generated using a uniform distribution over  $[0, 1]$ . The randomly chosen edge weights, denoted by  $a_l$ , where  $l \in \{1, 2, \dots, \frac{n^2-n}{2}\}$ , and without loss of generality assumed to be in increasing order, were assigned to edges using the following algorithm such that  $w_{i,j} < w_{i,j+1}$  and  $w_{i,j} < w_{i+1,j}$ .

**Step 1:** Initialize  $l \leftarrow 1$ ;

**Step 2:** for  $i = 1$  to  $n-1$

**Step 3:** for  $j = i+1$  to  $n$

**Step 3:**  $w_{ij} \leftarrow a_l$ ;

**Step 4:**  $l \leftarrow l+1$ ;

**Step 5:** end-for

**Step 6:** end-for

A correlative relationship among SNP variables, or linkage disequilibrium (LD), is the non-random association between SNP alleles. The distribution of these relationships among SNPs in a given genome tends to be greater when SNPs are closely located; this correlation diminishes quickly as genomic distance between SNPs gets larger, e.g. LD decay [63]. As a result, the distribution of correlative relationships among SNPs is a mixture of a small number of highly correlated SNPs with a large number of SNPs in low correlations. Assigning edge weights in increasing order is a simple way to guarantee that only part of the vertices has low edge weights close to 0, which can be used to define the independent variables. Meanwhile, we can also identify a subset of vertices with edge weights higher than a predefined threshold within this set, which could be used to define the independent variables and highly correlated variables.

A vertex  $i$  that has all neighbors with  $w_{ij} < 0.1$ , where  $j \in V$  and  $j \neq i$ , was defined as an independent variable. The subset generated by SNP-SELECT has to include all the independent variables to confirm that the  $k$ -dominating set based approach is able to identify independent variables. Highly correlated variables were defined as a subset ( $P$ ) of the variables where  $P \subset V$ , and the edge weights ( $w_{ij}^P$ ) within this subset are greater than a predefined threshold. In this simulation, we selected 0.8, 0.6, 0.4, and 0.3 as the predefined thresholds for the purpose of

illustrating the capability of the proposed approach to select the highly correlated variables. If SNP-SELECT includes at least one of the predefined highly correlated variables, the performance of the algorithm in selecting proxy variables among the highly correlated ones is considered fulfilled.

### Douglas-fir breeding populations

The Douglas-fir breeding population was established by the Ministry of Forests, Lands and Natural Resource of British Columbia, Canada in 1975 and consists of 165 full-sib families generated from structured paired-matings among 54 parents. The 1,372 individual trees used in this study consist of a subset of the full population and contains 37 full-sib families from 38 parents (see [64], for complete details). SNP genotypes for these 1,372 trees were generated using exome capture [65], resulting in 106,099 SNPs with missing ratio threshold less than 25% and minimum minor allele frequency (MAF) greater than 5% which comprises the ‘original’ data set.

The average numerator relationship matrix (*A*-matrix) of the DF dataset is known due to the structured pedigree, and was used as a baseline for comparison. We calculated the genomic estimated relatedness (*G*-matrix) using R package “rrBLUP” [66] using the mean imputation option on the original SNP dataset, as well as the five *k*-dominating SNP subsets. The comparison of the pedigree-based relatedness (*A*-matrix) elements with those of the *G*-matrices of the selected SNP subsets was performed to validate that SNP-SELECT is able to minimize the deviation of diagonal elements while obtaining comparable genetic covariance among individuals (off-diagonal elements).

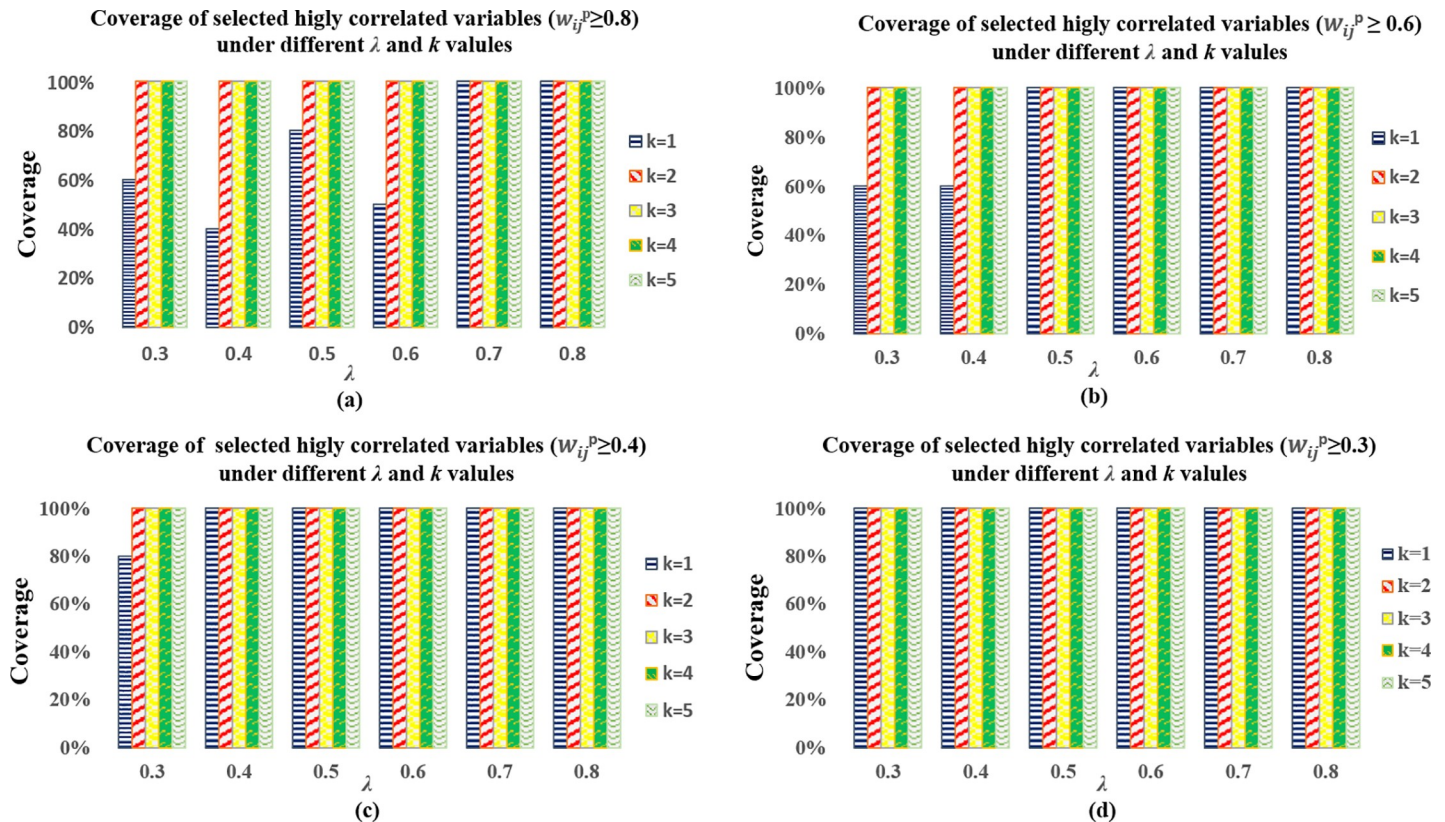
### Grasshopper mouse SNP data

Grasshopper mouse (genus, *Onychomys*) are cricetid rodents that inhabit prairies, deserts and desert grasslands throughout the western United States, northern Mexico, and south-central Canada [67]. Whereas *O. leucogaster* is readily distinguished based on body size, the two smaller species, *O. arenicola* and *O. torridus*, are morphologically cryptic and were treated as a single species until 1979 [68]. The SNP dataset analyzed here was generated using genotyping-by-sequencing, GBS [29], as part of a study designed to test for evidence of hybridization at a site in southwestern New Mexico where all three species come into contact [69], and at other sites in New Mexico and Arizona where *O. leucogaster* is sympatric with *O. arenicola* and *O. torridus*, respectively. SNPs were called using a reference-free SNP discovery protocol (UNEAK pipeline [70]), and filtered with minor allele frequency greater than 5% and missing ratio less than 10%.

## Results

### Simulation studies

When the SNP-SELECT algorithm was applied to the synthetic network with  $k \in \{1, 2, 3, 4, 5\}$ , all the *k*-dominating sets found included the predefined independent variables. The performance of the *k*-dominating set model in the selection of proxy variables is presented in Fig 2, with the predefined highly correlated variable thresholds  $\lambda \in \{0.8, 0.7, 0.6, 0.5, 0.4, 0.3\}$ . As shown in Fig 2(A), the definition of highly correlated variables was strict ( $w_{ij}^p > 0.8$ ); under this condition of few, highly connected variables, the use of larger values of either *k* or  $\lambda$  was encouraged. Also shown in Fig 2(B), 2(C) and 2(D), when relationships between variables are a mixture of high and low correlations, our results suggest the use of smaller values in *k* and  $\lambda$  to



**Fig 2. The capability of  $k$ -dominating set in selecting proxy variables among highly correlated variables.** Ten synthetic undirected networks with  $n = 1,000$  vertices ( $V$ ) were simulated. (a) highly correlated variables defined as  $w_{ij}^p > 0.8$ ; (b) highly correlated variables defined as  $w_{ij}^p > 0.6$ ; (c) highly correlated variables defined as  $w_{ij}^p > 0.4$ ; (d) highly correlated variables defined as  $w_{ij}^p > 0.3$ .

<https://doi.org/10.1371/journal.pone.0203242.g002>

capture all relationships. By varying on  $k$  and  $\lambda$ , we demonstrate the flexibility and strength of SNP-SELECT in choosing proxy variables from highly correlated variables.

### Pedigree recovery for Douglas-fir breeding populations

The SNP-SELECT algorithm was applied to select the influential SNPs to reconstruct the known pedigree for a Douglas-fir (DF) breeding population. Four  $k$ -dominating sets (DF107, DF105, DF103, DF102) with  $k = 1$ , and  $\lambda \in \{0.7, 0.5, 0.3, 0.2\}$  were generated. Among the four 1-dominating sets, DF103 has the best performance as shown in Table 1. To further investigate the impact of  $k$  on variable selection, another  $k$ -dominating set, DF203, with  $k = 2$  and  $\lambda = 0.3$  was generated to compare with DF103. The number of selected SNPs in DF107, DF105, DF103, DF102 and DF203 is 80,735, 67,062, 51,415, 41,539, and 68,188, respectively.

Without variable selection, original SNP data generated an average of 18% discrepancy on the diagonal elements. The performance of the five  $k$ -dominating subsets was showed in the reduced diagonal differences from the genomic relationship matrix ( $G$ -matrix) to the traditional pedigree-based average numerator relationship matrix ( $A$ -matrix) (Table 1). Comparing the five  $k$ -dominating sets indicated that the DF103 subset performed best on pedigree reco, especially for the diagonal pedigree information recovery. Fig 3 further illustrates the efficiency of the DF103 subset on pedigree reconstruction, and indicates that the  $G$ -matrix generated from the DF103 subset was closer to the known  $A$ -matrix as compared with the original dataset's  $G$ -matrix. Additionally, we randomly selected 10 subsets with the same SNP number as



**Table 1. The average difference of the upper triangle and the diagonal between pedigree-based relatedness (*A*-matrix) and genomic estimated relatedness (*G*-matrix). The best selected-subset for pedigree reconstruction (subset DF103) is highlighted.  $\lambda$  is linkage disequilibrium estimate.**

|               | <i>k</i> | $\lambda$  | Num. of SNP   | Ave. difference upper triangle | Ave. difference diagonal |
|---------------|----------|------------|---------------|--------------------------------|--------------------------|
| Original Data | -        | -          | 106,099       | 0.034353                       | 0.180374                 |
| DF107         | 1        | 0.7        | 80,735        | 0.034240                       | 0.103673                 |
| DF105         | 1        | 0.5        | 67,062        | 0.034139                       | 0.055994                 |
| <b>DF103</b>  | <b>1</b> | <b>0.3</b> | <b>51,415</b> | <b>0.034018</b>                | <b>0.019769</b>          |
| DF102         | 1        | 0.2        | 41,539        | 0.034249                       | 0.123494                 |
| DF203         | 2        | 0.3        | 68,188        | 0.034180                       | 0.123950                 |
| Random subset | -        | -          | 51,415        | 0.034498                       | 0.180419                 |
| COR03         | -        | 0.3        | 39,768        | 0.034774                       | 0.234326                 |
| LRTag03       | -        | 0.3        | 51,022        | 0.034292                       | 0.135324                 |

<https://doi.org/10.1371/journal.pone.0203242.t001>

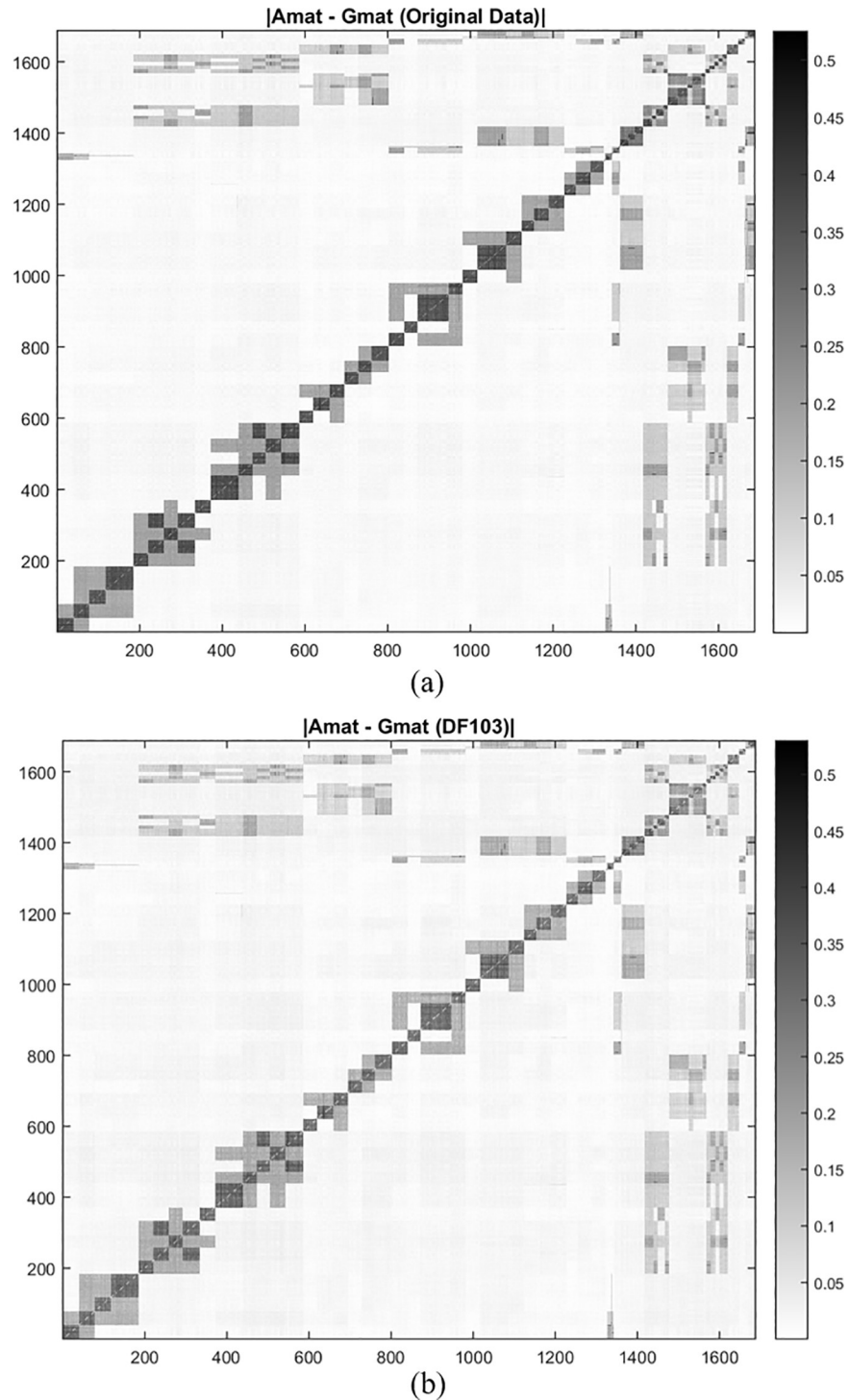
DF103 from the original dataset and used the average results of these 10 subsets to represent the performance of the randomly selected subset. The results indicated that all five *k*-dominating sets outperformed the randomly selected subset (Table 1).

The effectiveness of SNP-SELECT was also examined by the conventional approach that filters for SNP variables by pairwise correlation coefficients, as well as the LRTag algorithm that applies minimum set covering for SNP selection [71]. The discrepancy between *A*- and *G*-matrices resulted from using correlation coefficient of 0.3 and  $\lambda = 0.3$  was listed in Table 1 as COR03 and LRTag03, for pairwise correlation coefficient method and LRTag algorithm, respectively. Among all tests, the DF103 from SNP-SELECT remained the best SNP subset for estimating genetic relationship of Douglas-fir breeding population. Consider computing time requirement, when values of distance or pair-wise linkage disequilibria were pre-computed, SNP variable selection for SNP-SELECT could be complete in 8–10 minutes, while LRTag required about 18 hours for the same datasets.

### Clustering analysis for grasshopper mouse populations

To investigate parameters influencing population genetics of grasshopper mouse populations, 85,812 SNPs were used to genotype 226 samples representing three species: *O. arenicola* (*n* = 76), *O. leucogaster* (*n* = 67), and *O. torridus* (*n* = 83), collected at 12 geographic locations (Table 2). The dataset was pre-filtered based on a maximum of 10% missing data, and minimum MAF (minor allele frequency) of 5%. The SNP-SELECT was applied to generate three SNP subsets (MICE103, MICE105 and MICE107) with *k* = 1 and  $\lambda \in \{0.3, 0.5, 0.7\}$ , respectively; the number of informative SNPs retained in MICE103, MICE105 and MICE107 was 2,144, 11,014, and 22,355, respectively. The missing data in the original dataset and the three *k*-dominating sets was imputed with the most frequent genotype. Before the geographic origin analysis, we split the 226 samples into 3 groups based on species identity. There were 5 sampling locations in each species group (Table 2).

The performance of the three *k*-dominating sets' ability to predict the geographic origin of samples within each species was first evaluated using the *k*-means clustering approach in R [72]. Clustering was initiated with *k* = 5, random seed at 20 and *nstart* = 100, where *nstart* specifies the initial configurations, and the algorithm will report on the best one [73, 74]. The Adjusted Rand Index (ARI), a measure of agreement between clustering results and external criteria [75, 76], was used to evaluate the clustering results. As shown in Table 3, the clustering results for the largest SNP subset, MICE107, had the same performance as the original data of



**Fig 3.** (a) Heatmap of the absolute difference between pedigree-based relatedness (*A*-matrix) and genomic estimated relatedness (*G*-matrix) generated from original data; (b) Heatmap of the absolute difference between pedigree-based relatedness (*A*-matrix) and genomic estimated relatedness (*G*-matrix) generated from DF103 subset. The color of Fig 3 (B) is lighter than Fig 3(A). The lighter the color, the closer the relationship between *A*- and *G*-matrices of Douglas-fir breeding population.

<https://doi.org/10.1371/journal.pone.0203242.g003>

**Table 2. Geographic location of grasshopper mouse (*Onychomys*) samples.**

| Species               | Site Name (Sample Size)   |
|-----------------------|---|
| <i>O. arenicola</i>   | Animas/Rodeo, NM (20);<br>Pancho Villa, Chihuahua, Mex (7);<br>Organ Mountains, NM (27);<br>Sevilleta National Wildlife Refuge, NM (14);<br>Hidalgo del Parral, Chihuahua, Mex (8). |
| <i>O. leucogaster</i> | Petrified Forest, AZ (13);<br>Animas/Rodeo, NM (11);<br>Sevilleta National Wildlife Refuge, NM (14);<br>Felt, OK (19);<br>Garden City, KS (10).                                     |
| <i>O. torridus</i>    | Lone Pine, CA (11);<br>Carefree, AZ (8);<br>Santa Rita Experimental Range, AZ (19);<br>Animas/Rodeo, NM (28);<br>Chiricahua Mountains, AZ (17).                                     |

<https://doi.org/10.1371/journal.pone.0203242.t002>

85,812 SNPs in recovering the geographic origin of *O. arenicola* and *O. torridus* samples; however, MICE107 subset outperformed the original SNP data in recovering the geographic origin of *O. leucogaster* samples. Moreover, the clusters resulting from MICE105 and MICE103 exhibited larger ARI values than those from the original SNP data, indicative of a greater agreement and reduced errors in the clustering reached by SNP-SELECT variable selection. Overall, the MICE105 SNP subset (11,014 SNPs) demonstrated the greatest agreement among all selected subsets (Table 3).

To confirm that the performance of SNP-SELECT was not the result of a specific clustering algorithm, the partitioning around medoids (PAM) algorithm [77] with  $k = 5$  (random seed = 20) was performed using samples' dissimilarity matrix of each species. To describe the

dissimilarity matrix, we first define the  $\mathbf{G}$ -matrix as 
$$\begin{bmatrix} g_{11} & \dots & g_{1n} \\ \vdots & \dots & \vdots \\ g_{n1} & \dots & g_{nn} \end{bmatrix}$$
. Then the dissimilarity

matrix is defined as 
$$\begin{bmatrix} 1 - |g_{11}| & \dots & 1 - |g_{1n}| \\ \vdots & \dots & \vdots \\ 1 - |g_{n1}| & \dots & 1 - |g_{nn}| \end{bmatrix}$$
. In Table 3, clusters resulting from the PAM

algorithm also demonstrated that the selected subsets perform better than the original data in predicting actual sampling localities.

**Table 3. The adjusted rand index (ARI) shows the agreement between the computed clusters using  $k$ -means clustering algorithm and partitioning around medoids (PAM) algorithm with  $k = 5$ , using the original grasshopper mouse SNP data set and the  $k$ -dominating subsets. ARI values listed below show the agreement measurement between original sample locations and clustering results.**

| Method     | Dataset        | SNP Number    | <i>O. arenicola</i> | <i>O. leucogaster</i> | <i>O. torridus</i> |
|------------|----------------|---------------|---------------------|-----------------------|--------------------|
| $k$ -means | Original data  | 85,812        | 0.3868              | 0.5981                | 0.5692             |
|            | MICE107        | 22,355        | 0.3868              | 0.7158                | 0.5692             |
|            | <b>MICE105</b> | <b>11,014</b> | <b>0.5256</b>       | <b>0.7158</b>         | <b>0.6003</b>      |
|            | MICE103        | 2,144         | 0.3963              | 0.7158                | 0.6003             |
| PAM        | Original data  | 85,812        | 0.0706              | 0.2229                | 0.2244             |
|            | <b>MICE107</b> | 22,355        | 0.0513              | <b>0.3852</b>         | 0.1812             |
|            | MICE105        | 11,014        | 0.1016              | 0.3509                | 0.2445             |
|            | <b>MICE103</b> | 2,144         | <b>0.1172</b>       | 0.2902                | <b>0.2793</b>      |

<https://doi.org/10.1371/journal.pone.0203242.t003>

## Discussion

Owing to technological advancement in DNA sequencing methods, life scientists are grappling with exceedingly large data sets [78]. The most obvious challenge is the large amount of genomic variation that needs to be processed and quantified in a very short time period [79]. Although various data techniques have been adopted, the resulting data sets have several characteristics that make downstream analyses challenging [80]. The common ones are: the number of variables is often much larger than the number of individuals, and data sets are usually sparse regarding relevant information, *i.e.* only a small subset of variables is associated with the phenotypic variation [81].

In genetic analyses using high dimensional data sets where there are more parameters than observations, penalized regression techniques are often required to ensure stable estimates [82, 83]. The estimates of SNP marker effects are strongly affected by collinearity between predictors through a “grouping effect”- groups of variables highly correlated with other groups (of variables) sporadically [84]. Such multicollinearity would further confound gene expression values obtained from RNAseq or determination of significance of SNP causality in genome-wide association (GWAS) or genomic selection (GS) studies [85–87]. As a result, multiple-step GWAS and GS analysis that includes SNP variable selection has been explored [26, 88–91]. While adoption of these methods might be an advantage when seeking functional variants associated with traits of interest, these fitness-associated SNP variables would bias inferences of gene flow, migration or dispersal [37, 92], and estimates of relatedness and inbreeding depression [93].

Without the dependency on phenotypes, SNP variable selection methods currently focus on pairwise correlations between variables (e.g. [94]). In principle, SNP variables are selected if the absolute value of a pairwise correlation ( $|corr(i,j)|$ ) is less than a predefined threshold  $\lambda$ ; or if  $|corr(i,j)|$  is no less than the given threshold, only the second variable will be selected (e.g. if  $|corr(i,j)| \geq \lambda$ , SNP  $j$  will be selected). Here, we demonstrate the superior performance of the proposed  $k$ -dominating set variable selection over the conventional method of pairwise correlation coefficients (Table 1, COR03). As shown in Fig 3, diagonal values, indicative of the errors in estimating individuals’ genomic relationship based on markers, were minimized using SNP-SELECT. The pairwise estimates of genomic relationships (off-diagonal elements) were, however, mostly preserved (Table 1), suggesting that both the hidden and historical relatedness among individuals could still be recovered by the set of SNP variables selected by SNP-SELECT.

The use of genomic markers to uncover hidden relationships and potential pedigree in open-pollinated progeny has been effective in tree breeding programs [95, 96]. Such pedigree reconstruction is a preferred method to determine the genealogical relationship among groups of related individuals, leading to improved estimation of genetic parameters [97–99]. To maximize the advantage of using dense genomic markers, VanRaden [100] derived estimates of marker-based relationships between pairs of individuals as a genomic relationship matrix (**G**-matrix), which can be used as a substitute for the traditional pedigree-based average numerator relationship matrix (**A**-matrix) in Henderson’s animal model [101–103]. Also, combining the **A**-matrix and the **G**-matrix into a single genetic relationship matrix (**H**-matrix) has proven to be an effective approach to improve the relationship coefficients for better genetic parameter estimation [104, 105] and marker effect estimation [106], and to leverage extra phenotypic information from the non-genotyped individuals [103]. To ensure improved accuracy in such single-step methods, the **G**- and **A**-matrices should be compatible [107], and diagonal elements in the **G**-matrix need to be consistent with the **A**-matrix diagonal elements; therefore rescaling **A**- and **G**-matrices would reflect the mean difference between these matrices [108], a context in which using SNP markers selected by SNP-SELECT could be considerably beneficial.

## Conclusions

The  $k$ -dominating set model provides a flexible and effective method for selecting informative SNPs; a C++ source code (SNP-SELECT) that uses Gurobi™ Optimization solver is also released with the manuscript. This approach is scalable through the use of integer programming solvers and graph preprocessing, and can be extended to other genomic applications.

Using pedigree reconstruction and cluster analysis, the capacity of SNP-SELECT was demonstrated for solving the variable selection conundrum of large datasets without any significant runtime considerations. Furthermore, SNP-SELECT does not depend on the use of LD to define threshold for edges; other similarity/distance measure would broaden its applicability beyond breeding science and ecological genetics. Future work on the algorithmic aspects of this approach could focus on the development of graph and model decomposition techniques, as well as preprocessing techniques to improve scalability in practice.

## Acknowledgments

We thank Michael Stoebr from the Ministry of Forests, Lands and Natural Resource Operations, British Columbia, for the use of Douglas-fir data. We are also grateful to Ashlee Rowe, and to the following institutions for providing grasshopper mouse tissue samples: Museum of Southwestern Biology, Museum of Vertebrate Zoology, Oklahoma State University Collection of Vertebrates. The project was completed with support from the High Performance Computing Center Facilities of Oklahoma State University.

## Author Contributions

**Conceptualization:** Zhuqi Miao, Balabhaskar Balasundaram, Charles Chen.

**Data curation:** Shuzhen Sun, Blaise Ratcliffe, Polly Campbell, Bret Pasch, Yousry A. El-Kassaby.

**Formal analysis:** Shuzhen Sun, Zhuqi Miao, Charles Chen.

**Funding acquisition:** Charles Chen.

**Investigation:** Charles Chen.

**Methodology:** Shuzhen Sun, Balabhaskar Balasundaram, Charles Chen.

**Resources:** Charles Chen.

**Software:** Shuzhen Sun.

**Supervision:** Balabhaskar Balasundaram, Charles Chen.

**Writing – original draft:** Shuzhen Sun, Zhuqi Miao, Balabhaskar Balasundaram, Charles Chen.

**Writing – review & editing:** Shuzhen Sun, Blaise Ratcliffe, Polly Campbell, Bret Pasch, Yousry A. El-Kassaby, Balabhaskar Balasundaram, Charles Chen.

## References

1. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*. 2010; 20(1):101. PMID: [21572976](https://pubmed.ncbi.nlm.nih.gov/21572976/)
2. Hall P, Pittelkow Y, Ghosh M. Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2008; 70(1):159–173.

3. Kirpich A, Ainsworth EA, Wedow JM, Newman JRB, Michailidis G, McIntyre LM. Variable selection in omics data: A practical evaluation of small sample sizes. *PloS one*. 2018; 13(6):e0197910–e. <https://doi.org/10.1371/journal.pone.0197910> PMID: 29927942
4. Fan J, Han F, Liu H. Challenges of Big Data Analysis. *National science review*. 2014; 1(2):293–314. <https://doi.org/10.1093/nsr/nwt032> PMID: 25419469
5. Bakker MG, Manter DK, Sheflin AM, Weir TL, Vivanco JM. Harnessing the rhizosphere microbiome through plant breeding and agricultural management. *Plant and Soil*. 2012; 360(1):1–13.
6. Fan J, Guo S, Hao N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society Series B, Statistical methodology*. 2012; 74(1):37–65. <https://doi.org/10.1111/j.1467-9868.2011.01005.x> PMID: 22312234
7. Heinze G, Wallisch C, Dunkler D. Variable selection—A review and recommendations for the practicing statistician. *Biometrical journal Biometrische Zeitschrift*. 2018; 60(3):431–449. <https://doi.org/10.1002/bimj.201700067> PMID: 29292533
8. Zhang M, Zhang D, Wells MT. Variable selection for large p small n regression models with incomplete data: mapping QTL with epistases. *BMC bioinformatics*. 2008; 9:251. <https://doi.org/10.1186/1471-2105-9-251> PMID: 18510743
9. Lynch M, Xu S, Maruki T, Jiang X, Pfaffelhuber P, Haubold B. Genome-wide linkage-disequilibrium profiles from single individuals. *Genetics*. 2014; 198(1):269–281. <https://doi.org/10.1534/genetics.114.166843> PMID: 24948778
10. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. *Nature*. 2001; 411(6834):199–204. PMID: 11346797
11. Long N, Gianola D, Rosa GJM, Weigel KA. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. *Journal of Animal Breeding and Genetics*. 2011; 128(4):247–257. <https://doi.org/10.1111/j.1439-0388.2011.00917.x> PMID: 21749471
12. Song J, Carver BF, Powers C, Yan L, Klápště J, El-Kassaby YA, et al. Practical application of genomic selection in a doubled-haploid winter wheat breeding program. *Mol Breed*. 2017; 37(10):117. <https://doi.org/10.1007/s11032-017-0715-8> PMID: 28936114
13. Long N, Gianola D, Rosa GJ, Weigel KA, Avendano S. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *Journal of Animal Breeding and Genetics*. 2007; 124(6):377–389. <https://doi.org/10.1111/j.1439-0388.2007.00694.x> PMID: 18076475
14. Habier D, Fernando RL, Dekkers JC. Genomic selection using low-density marker panels. *Genetics*. 2009; 182(1):343–353. <https://doi.org/10.1534/genetics.108.100289> PMID: 19299339
15. Usai MG, Goddard ME, Hayes BJ. LASSO with cross-validation for genomic selection. *Genet Res (Camb)*. 2009; 91(6):427–436.
16. Song J, Carver BF, Powers C, Yan L, Klápště J, El-Kassaby YA, et al. Practical application of genomic selection in a doubled-haploid winter wheat breeding program. *Molecular Breeding*. 2017.
17. Weigel KA, de los Campos G, Gonzalez-Recio O, Naya H, Wu XL, Long N, et al. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J Dairy Sci*. 2009; 92(10):5248–57. <https://doi.org/10.3168/jds.2009-2092> PMID: 19762843
18. Pes B, Dessì N, Angioni M. Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data. *Information Fusion*. 2017; 35:132–147.
19. He Q, Lin D-Y. A variable selection method for genome-wide association studies. *Biometrics*. 2011; 27(1):1–8.
20. Ayers KL, Cordell HJ. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiol*. 2010; 34(8):879–91. <https://doi.org/10.1002/gepi.20543> PMID: 21104890
21. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol*. 1996; 58:267–288.
22. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in Partial Least Squares Regression. *Chemometrics Intellig Lab Syst*. 2012; 118:62–69.
23. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J Roy Stat Soc Ser B (Stat Method)*. 2008; 70(5):849–911.
24. Wasserman L, Roeder K. High dimensional variable selection. *Annals of statistics*. 2009; 37(5A):2178–201. PMID: 19784398
25. Bogdan M, van den Berg E, Sabatti C, Su W, Candès EJ. SLOPE—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*. 2015; 9(3):1103–1140. <https://doi.org/10.1214/15-AOAS842> PMID: 26709357

26. Dehman A, Ambroise C, Neuvial P. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*. 2015; 16(1):148.
27. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet*. 2003; 4:981. <https://doi.org/10.1038/nrg1226> PMID: 14631358
28. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet*. 2010; 6(2): e1000862. <https://doi.org/10.1371/journal.pgen.1000862> PMID: 20195501
29. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. 2011; 6(5):e19379. <https://doi.org/10.1371/journal.pone.0019379> PMID: 21573248
30. Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. *Tree Genet Genom*. 2013; 9(6):1537–1544.
31. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, et al. Detecting selection in population trees: The lewontin and krakauer test extended. *Genetics*. 2010; 186(1):241–262. <https://doi.org/10.1534/genetics.104.117275> PMID: 20855576
32. Beaumont MA, Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol*. 2004; 13(4):969–980. PMID: 15012769
33. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics*. 2008; 180(2):977–993. <https://doi.org/10.1534/genetics.108.092221> PMID: 18780740
34. Guo F, Dey DK, Holsinger KE. A bayesian hierarchical model for analysis of Single-Nucleotide Polymorphisms diversity in multilocus, multipopulation samples. *Journal of the American Statistical Association*. 2009; 104(485):142–154. <https://doi.org/10.1198/jasa.2009.0010> PMID: 19623271
35. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. *Annu Rev Genet*. 2013; 47(1):97–120.
36. Nielsen R. Statistical tests of selective neutrality in the age of genomics. *Heredity*. 2001; 86:641. PMID: 11595044
37. Kirk H, Freeland JR. Applications and implications of neutral versus non-neutral markers in Molecular Ecology. *Int J Mol Sci*. 2011; 12(6):3966. <https://doi.org/10.3390/ijms12063966> PMID: 21747718
38. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. *PLoS Genet*. 2013; 9(10):e1003905. <https://doi.org/10.1371/journal.pgen.1003905> PMID: 24204310
39. Robertson A. Gene frequency distributions as a test of selective neutrality. *Genetics*. 1975; 81(4):775–785. PMID: 1213275
40. Jain AK, Dubes RC. Algorithms for clustering data: Prentice-Hall, Inc.; 1988.
41. Jambu M, Lebeaux M-O. Cluster analysis and data analysis: Elsevier Science Ltd; 1983.
42. Spath H. Cluster analysis algorithms for data reduction and classification of objects: Chichester: Ellis Horwood; 1980.
43. West DB. Introduction to graph theory: Prentice hall Upper Saddle River; 2001.
44. Diestel R. Graph Theory. 5 ed: Springer-Verlag Berlin Heidelberg; 2017. XVIII, 429 p. <https://doi.org/10.1002/jgt.22225>
45. Haynes TW, Hedetniemi S, Slater P. Fundamentals of domination in graphs: Marcel Dekker Inc.; 1998.
46. White JG, Southgate E, Thomson JN, Brenner S. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society of London Series B*. 1986; 314:1–340. <https://doi.org/10.1098/rstb.1986.0056> PMID: 22462104
47. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature*. 1998; 393:440. <https://doi.org/10.1038/30918> PMID: 9623998
48. Balasundaram B, Butenko S. Graph domination, coloring and cliques in telecommunications. In: Resende MGC, Pardalos PM, editors. *Handbook of Optimization in Telecommunications*. Boston, MA: Springer US; 2006. p. 865–890.
49. Michael RG, David SJ. Computers and intractability: a guide to the theory of NP-completeness: WH Free. Co., San Fr; 1979. 90–1 p.
50. Butenko S, Cheng X, Oliveira CA, Pardalos PM. A new heuristic for the minimum connected dominating set problem on ad hoc wireless networks. In: Butenko S, Murphey R, Pardalos PM, editors. *Recent Developments in Cooperative Control and Optimization*. Boston, MA: Springer US; 2004. p. 61–73.

51. Wolsey LA. Integer Programming. Wiley; 1998.
52. Wang C, Kao W-H, Hsiao CK. Using hamming distance as information for SNP-sets clustering and testing in disease association studies. *PLoS One*. 2015; 10(8):e0135918. <https://doi.org/10.1371/journal.pone.0135918> PMID: 26302001
53. Bartlett CW, Yeon Cheong S, Hou L, Paquette J, Yee Lum P, Jäger G, et al. An eQTL biological data visualization challenge and approaches from the visualization community. *BMC Bioinformatics*. 2012; 13(8):S8.
54. Zhang X, Pan F, Xie Y, Zou F, Wang W, editors. COE: a general approach for efficient genome-wide two-locus epistasis test in disease association study2009; Berlin, Heidelberg: Springer Berlin Heidelberg.
55. vonHoldt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, et al. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature*. 2010; 464:898. <https://doi.org/10.1038/nature08837> PMID: 20237475
56. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, et al. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human Genomics*. 2004; 1(4):274. <https://doi.org/10.1186/1479-7364-1-4-274> PMID: 15588487
57. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010; 42:565. <https://doi.org/10.1038/ng.608> PMID: 20562875
58. Liu G, Wang Y, Wong L. FastTagger: an efficient algorithm for genome-wide tag SNP selection using multi-marker linkage disequilibrium. *BMC Bioinformatics*. 2010; 11(1):66.
59. Weng L, Macciardi F, Subramanian A, Guffanti G, Potkin SG, Yu Z, et al. SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics*. 2011; 12(1):99.
60. González-Martínez SC, Grivet D. Association mapping in plants. Oraguzie NC, Rikkerink EHA, Gardner SE, Silva HND, editors: Springer, New York, NY; 2009. ix-x p.
61. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*. 1995; 12(5):921–927. <https://doi.org/10.1093/oxfordjournals.molbev.a040269> PMID: 7476138
62. Gurobi Optimization I. Gurobi optimizer reference manual 2018 [Available from: <http://www.gurobi.com>]
63. Chen C, DeClerck G, Tian F, Spooner W, McCouch S, Buckler E. PICARA, an analytical pipeline providing probabilistic inference about a priori candidates genes underlying genome-wide association QTL in plants. *PLoS One*. 2012; 7(11):e46596. <https://doi.org/10.1371/journal.pone.0046596> PMID: 23144785
64. Thistlethwaite FR, Ratcliffe B, Klápště J, Porth I, Chen C, Stoehr MU, et al. Genomic prediction accuracies in space and time for height and wood density of Douglas-fir using exome capture as the genotyping platform. *BMC Genomics*. 2017; 18(1):930. <https://doi.org/10.1186/s12864-017-4258-5> PMID: 29197325
65. Neves LG, Davis JM, Barbazuk WB, Kirst M. Whole-exome targeted sequencing of the uncharacterized pine genome. *The Plant Journal*. 2013; 75(1):146–156. <https://doi.org/10.1111/tpj.12193> PMID: 23551702
66. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*. 2011; 4(3):250–5.
67. Hall ER. The mammals of North America. second ed: John Wiley and Sons, New York; 1981.
68. Hinesley LL. Systematics and distribution of two chromosome forms in the southern grasshopper mouse, genus *onychomys*. *J Mammal*. 1979; 60(1):117–128.
69. Sullivan RM, Hafner DJ, Yates TL. Genetics of a contact zone between three chromosomal forms of the grasshopper mouse (genus *onychomys*): A reassessment. *J Mammal*. 1986; 67(4):640–659.
70. Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, et al. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet*. 2013; 9(1):e1003215. <https://doi.org/10.1371/journal.pgen.1003215> PMID: 23349638
71. Liu L, Wu Y, Lonardi S, Jiang T. Efficient genome-wide TagSNP selection across populations via the linkage disequilibrium criterion. *J Comput Biol*. 2010; 17(1):21–37. <https://doi.org/10.1089/cmb.2007.0228> PMID: 20078395
72. Team RDC. R: A Language and Environment for Statistical Computing. 2011.
73. Muca M, Kutrolli G, Kutrolli M. A proposed algorithm for determining the optimal number of clusters. *European Scientific Journal, ESJ*. 2015; 11(36).



74. Jay JJ, Eblen JD, Zhang Y, Benson M, Perkins AD, Saxton AM, et al. A systematic comparison of genome-scale clustering algorithms. *BMC Bioinformatics*. 2012; 13(10):S7.
75. Yeung KY, Ruzzo W. Details of the adjusted rand index and clustering algorithms supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data" (to appear in *Bioinformatics*)2001.
76. Santos JM, Embrechts M, editors. *On the use of the adjusted rand index as a metric for evaluating supervised classification*2009; Berlin, Heidelberg: Springer Berlin Heidelberg.
77. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. *cluster: cluster analysis basics and extensions*. R package version 2.0.1. 2015.
78. Marx V. The big challenges of big data. *Nature*. 2013; 498:255. <https://doi.org/10.1038/498255a> PMID: 23765498
79. May M. Life science technologies: big biological impacts from big data. *Science*. 2014; 344(6189):1298–1300.
80. Li Y, Chen L. Big biological data: challenges and opportunities. *Genomics Proteomics Bioinformatics*. 2014; 12(5):187–9. <https://doi.org/10.1016/j.gpb.2014.10.001> PMID: 25462151
81. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform*. 2017:bbx124-bbx.
82. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001; 157(4):1819–29. PMID: 11290733
83. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*. Springer series in statistics New York; 2001.
84. Wimmer V, Lehermeier C, Albrecht T, Auinger H-J, Wang Y, Schön C-C. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*. 2013; 195(2):573–587. <https://doi.org/10.1534/genetics.113.150078> PMID: 23934883
85. Hong S, Kim Y, Park T. Practical issues in screening and variable selection in genome-wide association analysis. *Cancer Inform*. 2014; 13(Suppl 7):55–65. <https://doi.org/10.4137/CIN.S16350> PMID: 25635166
86. Ishwaran H, Rao JS. Geometry and properties of generalized ridge regression in high dimensions. *Contemp Math*. 2014; 622:81–93.
87. El-Kassaby YA. Associations between allozyme genotypes and quantitative traits in Douglas-fir [*Pseudotsuga menziesii* (Mirb.) Franco]. *Genetics*. 1982; 101(1):103–115. PMID: 17246076
88. Cho S, Kim K, Kim YJ, Lee J-K, Cho YS, Lee J-Y, et al. Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Ann Hum Genet*. 2010; 74(5):416–428. <https://doi.org/10.1111/j.1469-1809.2010.00597.x> PMID: 20642809
89. Szymczak S, Holzinger E, Dasgupta A, Malley JD, Molloy AM, Mills JL, et al. r2VIM: a new variable selection method for random forests in genome-wide association studies. *BioData Min*. 2016; 9(1):7.
90. Meuwissen THE, Indahl UG, Ødegård J. Variable selection models for genomic selection using whole-genome sequence data and singular value decomposition. *Genetics, selection, evolution: GSE*. 2017; 49(1):94–. <https://doi.org/10.1186/s12711-017-0369-3> PMID: 29281962
91. Schulz-Streeck T, Ogutu JO, Piepho HP. Pre-selection of markers for genomic selection. *BMC proceedings*. 2011; 5 Suppl 3:S12.
92. Holderegger R, Kamm U, Gugerli F. Adaptive vs. neutral genetic diversity: implications for landscape genetics. *Landscape Ecol*. 2006; 21(6):797–807.
93. Chelo IM, Carvalho S, Roque M, Proulx SR, Teotónio H. The genetic basis and experimental evolution of inbreeding depression in *Caenorhabditis elegans*. *Heredity*. 2013; 112:248. <https://doi.org/10.1038/hdy.2013.100> PMID: 24129606
94. Hainke K, Szugat S, Fried R, Rahnenführer J. Variable selection for disease progression models: methods for oncogenetic trees and application to cancer and HIV. *BMC Bioinformatics*. 2017; 18(1):358. <https://doi.org/10.1186/s12859-017-1762-1> PMID: 28764644
95. Wang J. Sibship reconstruction from genetic data with typing errors. *Genetics*. 2004; 166(4):1963–1979. PMID: 15126412
96. Kalinowski ST, Taper ML, Marshall TC. Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Mol Ecol*. 2007; 16(5):1099–1106. <https://doi.org/10.1111/j.1365-294X.2007.03089.x> PMID: 17305863
97. El-Kassaby YA, LstibÚrek M. Breeding without breeding. *Genetics Research*. 2009; 91(2):111–120. <https://doi.org/10.1017/S001667230900007X> PMID: 19393127

98. Klápště J, Lstibůrek M, El-Kassaby YA. Estimates of genetic parameters and breeding values from western larch open-pollinated families using marker-based relationship. *Tree Genet Genom*. 2014; 10(2):241–249.
99. El-Kassaby YA, Cappa EP, Liewlaksaneeyanawin C, Klapste J, Lstiburek M. Breeding without breeding: is a complete pedigree necessary for efficient breeding? *PLoS One*. 2011; 6(10):e25737. <https://doi.org/10.1371/journal.pone.0025737> PMID: 21991342
100. VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci*. 2008; 91(11):4414–4423. <https://doi.org/10.3168/jds.2007-0980> PMID: 18946147
101. Henderson C. Applications of linear models in animal breeding. University of Guelph Press, Guelph. 1984; 11:652–653.
102. Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: a Look into the black box of genomic prediction. *Genetics*. 2013; 194(3):597–607. <https://doi.org/10.1534/genetics.113.152207> PMID: 23640517
103. Ratcliffe B, Gamal El-Dien O, Cappa EP, Porth I, Klapste J, Chen C, et al. Single-step BLUP with varying genotyping effort in open-pollinated *Picea glauca*. G3: Genes|Genomes|Genetics. 2017.
104. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. *J Dairy Sci*. 92(9):4656–63. <https://doi.org/10.3168/jds.2009-2061> PMID: 19700729
105. Misztal I, Legarra A, Aguilar I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J Dairy Sci*. 92(9):4648–55. <https://doi.org/10.3168/jds.2009-2064> PMID: 19700728
106. Wang H, Misztal I, Aguilar I, Legarra A, Muir WM. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genetics Research*. 2012; 94(2):73–83. <https://doi.org/10.1017/S0016672312000274> PMID: 22624567
107. Christensen OF, Madsen P, Nielsen B, Ostersen T, Su G. Single-step methods for genomic evaluation in pigs. *Animal*. 2012; 6(10):1565–71. <https://doi.org/10.1017/S1751731112000742> PMID: 22717310
108. Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*. 2010; 11:800. <https://doi.org/10.1038/nrg2865> PMID: 20877324