

A comparison of polymorphism information content and mean of transformed kinships as criteria for selecting informative subsets of barley (*Hordeum vulgare* L. s. l.) from the USDA Barley Core Collection

Ryan C. Graebner · Patrick M. Hayes ·
Christina H. Hagerty · Alfonso Cuesta-Marcos

Received: 13 January 2015 / Accepted: 5 May 2015 / Published online: 20 May 2015
© Springer Science+Business Media Dordrecht 2015

Abstract Recent advances in genetic technologies have given researchers the ability to characterize genetic marker data for large germplasm collections. While some studies are able to capitalize on entire germplasm collections, others, especially those that focus on traits that are difficult to phenotype, instead focus on a subset of the collection. Typically, subsets are selected using phenotypic or geographic data. One major hurdle in identifying favorable subsets is selecting a criterion that can be used to quantify the value of a subset. This study compares two such criteria, polymorphism information content, and a new criterion based on kinship matrices, which will be called the mean of transformed kinships. These criteria were explored in terms of their ability to select subsets that are favorable for genome wide association studies, and in their ability to select subsets that contain a high number of rare phenotypes. Using phenotypic and genotypic data that has been amassed from the USDA Barley Core Collection, evidence was found to support the hypotheses that subsets based on the mean of transformed kinships were well-suited to select subsets intended for genome-wide association studies, but

the same was not found for polymorphism information content. Inversely, evidence was found to support the hypothesis that subsets based on polymorphism information content were well-suited to select subsets intended for rare-phenotype discovery, but the same was not found for subsets selected using the mean of transformed kinships criterion. Tools to select subsets using these two criteria have been released in the R package “GeneticSubsetter.”

Keywords Barley · Germplasm · GWAS · *Hordeum vulgare* · MTK · PIC · Rare phenotypes · Subset

Introduction

Global efforts to preserve the genetic diversity of agriculturally important crops have resulted in a range of valuable germplasm collections. Projects screening germplasm collections for novel phenotypes and genes often do not have the resources to sample every accession in a given collection. Therefore, subsets of the total collections are made. Until recently, these subsets were generally made on the basis of phenotype and geographic origin of accessions, with the goal of maximizing genetic diversity (Holbrook et al. 1993; Mahajan et al. 1996; Upadhyaya et al. 2001, 2009; Zewdie et al. 2004). However, with the advent of high-throughput genotyping, complete sets of genotypic data are increasingly common for large germplasm collections (Muñoz-Amatráin et al. 2014). This

R. C. Graebner · P. M. Hayes (✉) · A. Cuesta-Marcos
Department of Crop and Soil Science, Oregon State
University, Crop Science Building, Corvallis, OR, USA
e-mail: patrick.m.hayes@oregonstate.edu

C. H. Hagerty
Department of Botany and Plant Pathology, Oregon State
University, Cordley Hall, Corvallis, OR, USA

enables researchers to directly observe genetic diversity, as opposed to estimating it with phenotypic or geographic information.

Two principal components to any subsetting technique are the criterion used to quantify the value of a specific subset, and the method used to find the optimum subset, as judged by that criterion. For smaller collections, the method used to identify a favorable subset could be to simply test all possible subsets. However, this quickly becomes unfeasible as population sizes grow. For instance, in a circumstance where 100 accessions need to be chosen from a collection of 1000 accessions, there could be 6.385×10^{139} possible subsets. Given the large number of subset combinations, alternative methods are needed to reach a satisfactory, or ideally the best, subset for a given criterion. Without proper subsetting techniques, important phenotypes could be omitted, making them unavailable to breeders.

To quantify a population's diversity, polymorphism information content (PIC) values were calculated with the following equation:

$$PIC = 1 - \frac{1}{m} \sum_{l=1}^m \sum_{i=1}^n f_{li}^2$$

where f_{li} is the frequency of the l th locus for m loci, and the i th allele for n alleles. This equation was modified from an equation described by Smith et al. (1997). This equation is also used to calculate heterozygosity, which in inbred accessions is generally used to describe the probability that two random accessions would have different alleles at a random locus. Generally speaking, for bi-allelic markers, mean PIC values for a population can range from 0, where all markers are monomorphic, to 0.5, where the frequency of both alleles is 0.5 for every marker. While PICs are most frequently used to quantify the diversity of an existing set of genotypes, they have also been used to identify informative subsets in the program PowerMarker (Liu and Muse 2005), and in a study characterizing the USDA Barley Core Collection (Muñoz-Amatrain et al. 2014). Because a complete description of the methods used by PowerMarker to identify subsets is no longer available, it will not be evaluated in this study.

One shortcoming of the PIC criterion is that it does a poor job at removing similar genotypes from a population, when the similar genotypes contain alleles

that are sufficiently rare in the population. To address this, an alternative approach has been developed based specifically on kinship matrices, where kinship values are risen to the power of ten in order to increase the weight of pairs of similar genotypes. Subsets are compared by simply comparing the mean of these modified kinship values, or the mean of transformed kinships (MTK).

Our objectives in this study were to assess the utility of these subsetting criteria, both in terms of their ability to select subsets that are favorable for genome-wide association studies (GWAS), and in their ability to select subsets that contain a high number of rare phenotypes. The functions used to identify favorable subsets in this study are available in the R package "GeneticSubsetter."

Materials and methods

Description of functions

To calculate the MTK for a set of genotypes, a kinship matrix was made using the "A.mat" function in the R package rrBLUP (Endelman 2011), using the default options. Due to the way the A.mat function calculates kinship matrices, negative kinship values are created, and the cell describing an accession's kinship with itself has a degree of variability. To remove negative values, the kinship matrix was scaled to values ranging from zero to two (where the relative distance between kinship values were constant, and zero and two were the lowest and highest kinship values for the particular set of genotypes, respectively). Due to the method used to calculate the kinship matrix by A.mat, diagonal values (the values describing an accession's kinship with itself) were not constant across the population. To avoid this inconsistency from becoming a major factor in which accessions were eliminated, these diagonal values were replaced with zero. Each value in the kinship matrix was raised to the power of ten to accentuate similarities between accessions. Finally, the mean of the values in the resulting transformed kinship matrix was calculated, to find MTK, which quantifies the extent to which a subset contains closely related accessions. To make this criterion computationally feasible for subsetting, transformed kinship values were calculated once using

the whole population. Then subsets of the matrix are used to calculate MTK for subsets of genotypes.

The core functions, “SubsetterPIC” and “SubsetterMTK,” in the R package GeneticSubsetter, remove one genotype at a time, on the basis of which genotype’s removal will result in the highest PIC, or the lowest MTK, respectively. These functions return a list of ranked genotypes, from which subsets of any size can be obtained by taking the top accessions. The SubsetterPIC function returns a list identical to the list returned by the Excel macro discussed in Muñoz-Amatraín et al. (2014). However, the SubsetterPIC function uses a more efficient algorithm to identify this ranking, giving it a considerable advantage in computing time over the Excel macro.

Currently, these functions are only designed for homozygous, bi-allelic markers. However, the concepts used to calculate PIC and MTK in these functions could be applied for heterozygous and poly-allelic markers.

Description of germplasm

Data from the USDA Barley Core Collection was used to test these subsetting criteria. This collection contains 2417 landraces, breeding lines, and cultivars that have been collected from around the world (Muñoz-Amatraín et al. 2014). This collection was selected from the larger National Small Grains Collection (NSGC) for barley, by randomly selecting accessions based on the logarithm of the total number of entries from each country of origin, ensuring that a minimum of one accession from each country be included in the core collection (Muñoz-Amatraín et al. 2014).

Analysis of effect on GWAS

The PIC and MTK criteria were evaluated using a hybrid data set, which consisted of real heading date data and genotypic data from the USDA Barley Core Collection, and was modified to include simulated quantitative trait loci (QTL). This allowed us to leverage the advantages of using simulated data (including control over the number and magnitude of QTL, and reduced ambiguity regarding the true effects of loci), with a realistic estimation of the effect population structure has on phenotypic data. To create the simulated QTL, 20 single nucleotide polymorphisms (SNPs) from the barley iSelect Illumina SNP

platform (Muñoz-Amatraín et al. 2014) were chosen at random, and the heading date data for one of the two genotypes was increased by 5 days. Subsets made using these criteria were assessed by their ability to identify simulated QTL using GWAS. Genotypes were ranked using the SubsetterPIC and SubsetterMTK, and 200 times randomly, to make a total of 202 set of nested subsets (SNSs). Each of these SNSs consisted of a series of subsets, one for each multiple of 50 genotypes between 150 and 1800 genotypes (a total of 35 subsets for each SNS), where each accession in a given subset was also present in the subsets that were larger than it in the given SNS. GWAS was conducted for each subset in each of the 202 SNSs. GWAS was conducted using the “GWAS” function in the R package rrBLUP, using the default parameters (Endelman 2011).

Within each subset size (of 35 subset sizes), SNSs were assigned a rank based on how many simulated QTL were detected, relative to subsets of that size within other SNSs. The mean of a SNS’s ranks across all 35 tested subset sizes was used to quantify a particular SNS’s performance against other SNSs. Simple methods for combining p values would not be appropriate here, as two subsets of a similar size from a single SNS are not independent from each other. While many random SNSs can be obtained from this collection, the SubsetterPIC and SubsetterMTK functions are determinate in nature, and were only able to return one SNS each. To test if a particular subsetting function returned a SNS that was better than a random SNS (with $p < 0.05$), the non-random SNS was compared to the 200 random SNSs. A non-random SNS performing either better or worse than 97.5 % of the random subsets would correspond to $p < 0.05$, in which case it would be decided that there was a significant difference between the SNS made using the particular criterion and the random SNSs, within the context of this collection.

Analysis of effect on rare phenotype discovery

Eleven extreme phenotypes were identified, where extreme phenotypes were defined as either the highest or lowest ~ 2 % of accessions for each given trait (Table 1). For example, the trait “plant height” had two sets of accessions that held an extreme phenotype: the 25 accessions that were shorter than 66 cm, and the 23 accessions that were taller than 117.5 cm. These

Table 1 Rare phenotypes used in this study, the criteria to define the rare phenotypes, and the number of accessions that fit these criteria in the set of 1099 genotypes used to compare the abilities of subsetting criteria in rare-phenotype discovery

Trait	Definition of phenotype	Rare phenotype frequency
Spot blotch resistance	<4 on a 1–9 scale	32
Russian wheat aphid resistance	<7 on a 1–9 scale	14
Early heading	<31 days after first heading date	18
Low plant height	<66 cm	25
High plant height	>117.5 cm	23
Low beta-glucan	<3.34 %	20
High beta-glucan	>7.04 %	20
Low protein	<9.075 %	20
High protein	>18.15 %	20
Low kernel weight	<31.75 mg	23
High kernel weight	>60.25 mg	19

extreme phenotypes were used to test whether these subsetting criteria were beneficial for the discovery of rare phenotypes. To circumvent the limitations of only having access to one large collection with extensive phenotypic and genotypic data available, we used 1099 genotypes with thorough phenotypic information available to make 1000 random “mini-sets” of 100 accessions. While these mini-sets have similar population structures, pairs of mini-sets share an average of only 9.1 % of their genotypes, making their results essentially independent from each other. Each mini-set was further subsetted three times to a subset size of ten genotypes, once using the SubsetterMTK function, once using the SubsetterPIC function, and once randomly. Each 10-genotype subset was quantified by how many of the original ten rare alleles were present in the final subset. Paired *t* tests were used to determine if either the SubsetterPIC or the SubsetterMTK functions were able to identify subsets with more rare alleles than randomly selected subsets.

Phenotypic and genotypic information

To test the PIC and MTK criteria, we used phenotypic and genotypic data collected from the USDA Barley Core Collection. The collection was previously genotyped, using a barley iSelect Illumina SNP platform, which included 7842 SNPs (Muñoz-Amatraín et al. 2014). Heading date data were collected in Corvallis, Oregon in 2012 (Muñoz-Amatraín et al. 2014). All other phenotypic data were collected from the USDA-ARS Germplasm Resources Information Network website. A total of 1852 accessions had complete heading date and genotypic data available. A total of

1099 accessions had complete information available for genotypic data, and each of the eleven rare phenotypes included in the analysis of these criteria’s effect on rare-trait discovery.

Results

Gwas

A SNS made using SubsetterMTK performed better than 199 out of the 200 random SNSs (Fig. 2). This corresponds to a *p* value of approximately 0.01, providing strong evidence that subsets made using the MTK criterion are more favorable for GWAS within the context of the USDA Barley Core Collection. A SNS made using SubsetterPIC performed better than 131 out of the 200 random SNSs, corresponding to a *p* value of approximately 0.69. While this presents no evidence that subsets made using the PIC criterion are more favorable for GWAS for this specific germplasm collection, given the extremely low power of this test, this criterion may still have a benefit to subsetting for GWAS that was undetectable in this study.

Rare phenotypes

We found significant evidence that subsets identified using the PIC criterion were more likely to contain rare phenotypes than random subsets in the USDA Barley Core Collection (*p* < 0.0001). However, we found no evidence that the same was true for subsets identified using the MTK criterion (*p* = 0.83). On average,

random subsets of ten genotypes from the USDA Barley Core Collection contained 1.97 rare or extreme phenotypes. In contrast, subsets of ten genotypes selected from 100 random genotypes using the SubsetterPIC function contained an average of 2.46 rare or extreme phenotypes, representing a 24 % increase over random subsetting.

Structure

Principal component analysis (PCA) plots showing the full collection, a completely random subset, a subset of

200 genotypes made using the PIC criterion, and a subset of 200 genotypes made using the MTK criterion (Fig. 1). These figures demonstrate how the PIC and MTK criteria differ in terms of the resulting population structure. While it appears that subsets made using SubsetterPIC maintain the general structure of the full collection, the number of individuals in each group appears to differ from the random subset. This is likely because the PIC criterion will weight groups by their contribution to the subset's diversity, while the random subset weights groups by purely by how well they are represented in the full collection. Using

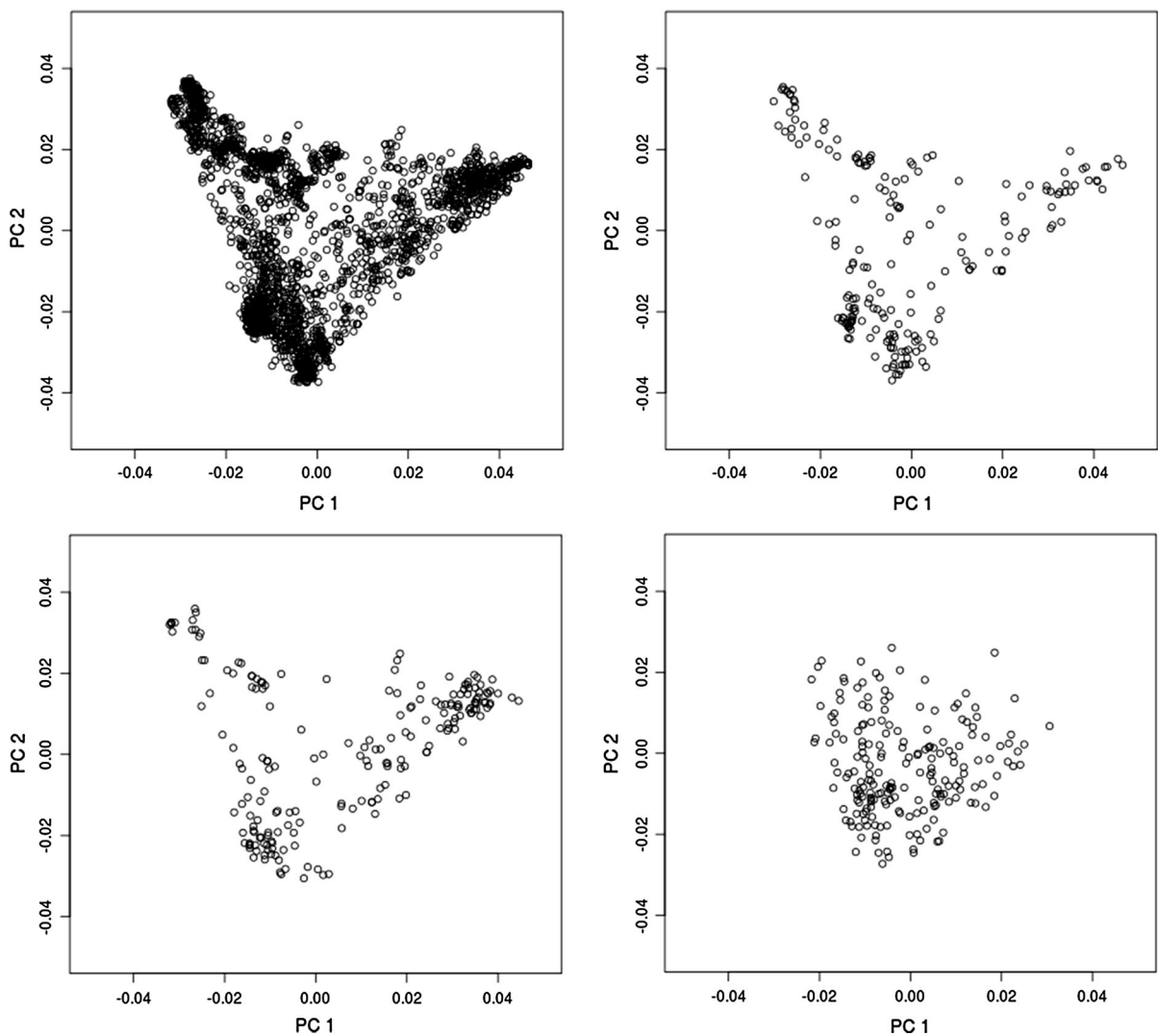


Fig. 1 PCA plots of the USDA Barley Core Collection (*top left*), a random subset (*top right*), a 200-genotype subset made using the SubsetterPIC function (*bottom left*), and a 200-genotype subset made using the SubsetterMTK function (*bottom right*)

SubsetterMTK instead appears to result in a population with very little structure. Interestingly, SubsetterMTK appears to prioritize genotypes that fall in the middle of the PCA plot, presumably because these genotypes are in fact the least related to the rest of the collection (Fig. 2).

Discussion

Within the context of the USDA Barley Core Collection, these results demonstrate that PIC is an acceptable subsetting criterion for rare phenotype discovery, and that MTK is an acceptable subsetting criterion for GWAS. Both functions were shown to avoid a loss of power when used to make subsets for their respective strengths. Due to the limited number of core collections that have been extensively phenotyped and genotyped, it is currently difficult to assess these benefits on other sets of accessions.

For dual-purpose subsets, it may be beneficial to use a combination of these two criteria (i.e. remove 100 accessions based on MTK, then an additional 100 accessions based on PIC). This approach may be able to maintain more than half of the benefit of only using one criteria, because these functions should first remove the accessions that contribute very little to

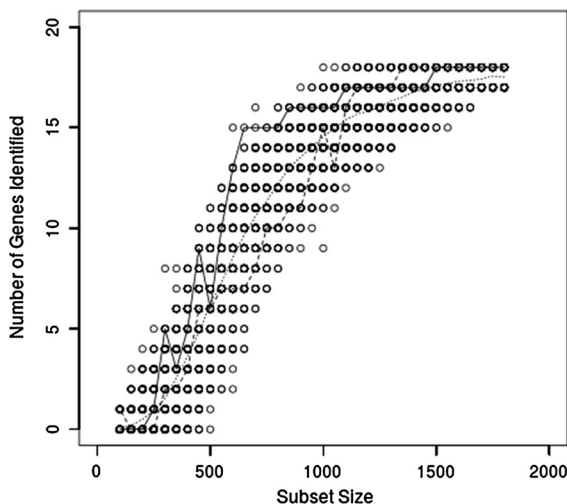


Fig. 2 Comparison of subsets identified by SubsetterMTK (solid line), SubsetterPIC (dashed line), and 200 subsets that were randomly selected (circles, with dotted line showing mean). The x-axis shows the size of each subset, and the y-axis show the number of artificial genes detected by that subset

the collections diversity, or that are essentially redundant, depending on the criteria used. Alternatively, a hybrid criterion could be used, which considers how each accession's removal would affect both the PIC and the MTK values for the subset.

These results suggest that the functions presented in the R package GeneticSubsetter can help to leverage “big data” in a way that substantially increases the efficiency of GWAS and rare-phenotype discovery: two tasks which are routinely conducted by plant breeding programs. While the R package GeneticSubsetter is currently only equipped to address homozygous accessions, we look forward to building on these functions to expand this package's utility to species that tend to be heterozygous, including humans and other animals.

Acknowledgments We would like to thank Professor Jennifer Kling and Dr. WTB Thomas for their guidance in developing this project.

References

- Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4:250–255
- Holbrook CC, Anderson WF, Pittman RN (1993) Selection of a core collection from the U.S. germplasm collection of peanut. *Crop Sci* 33:859–861
- Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21:2128–2129
- Mahajan RK, Bisht IS, Agrawal RC, Rana RS (1996) Studies on South Asian okra collection: methodology for establishing a representative core set using characterization data. *Genet Resour Crop Evol* 43:249–255
- Muñoz-Amatruán M, Cuesta-Marcos A, Endelman JB, Comadran J, Bonman JM (2014) The USDA barley core collection: genetic diversity, population structure, and potential for genome-wide association studies. *PLoS One* 9:e94688
- Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ et al (1997) An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparisons with data from RFLPS and pedigree. *Theor Appl Genet* 95:163–173
- Upadhyaya HD, Bramel PJ, Singh S (2001) Development of a chickpea core subset using geographic distribution and quantitative traits. *Crop Sci* 41:206–210
- Upadhyaya HD, Pundir RPS, Dwivedi SL, Gowda CLL, Reddy VG, Singh S (2009) Developing a mini core collection of sorghum for diversified utilization of germplasm. *Crop Sci* 49:1769–1780
- Zewdie Y, Tong N, Bosland P (2004) Establishing a core collection of Capsicum using a cluster analysis with enlightened selection of accessions. *Genet Resour Crop Evol* 51:147–151