Theor Appl Genet (2013) 126:379–388
DOI 10.1007/s00122-012-1986-4

ORIGINAL PAPER

Selecting subsets of genotyped experimental populations for phenotyping to maximize genetic diversity

B. Emma Huang · David Clifford · Colin Cavanagh

Received: 21 May 2012 / Accepted: 15 September 2012 / Published online: 5 October 2012

Abstract Selective phenotyping is a way of capturing the benefits of large population sizes without the need to carry out large-scale phenotyping and hence is a cost-effective means of capturing information about gene–trait relationships within a population. The diversity within the sample gives an indication of the efficiency of this information capture; less diversity implies greater redundancy of the genetic information. Here, we propose a method to maximize genetic diversity within the selected samples. Our method is applicable to general experimental designs and robust to common problems such as missing data and dominant markers. In particular, we discuss its application to multi-parent advanced generation intercross (MAGIC) populations, where, although thousands of lines may be genotyped as a large population resource, only hundreds may need to be phenotyped for individual studies. Through simulation, we compare our method to simple random sampling and the minimum moment aberration method. While the gain in power over simple random sampling for all tested methods is not large, our method results in a much more diverse sample of genotypes. This diversity can be applied to improve fine mapping resolution once a QTL region has been detected. Further, when applied to two wheat datasets from doubled haploid and MAGIC progeny, our method detects known QTL for small sample sizes where other methods fail.

Communicated by I. Mackay.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-1986-4) contains supplementary material, which is available to authorized users.

B. Emma Huang (✉)
CSIRO Mathematics, Informatics and Statistics and Food Futures Flagship, Dutton Park, QLD, Australia
e-mail: emma.huang@csiro.au

D. Clifford
CSIRO Mathematics, Informatics and Statistics, Dutton Park, QLD, Australia

C. Cavanagh
CSIRO Plant Industry, Black Mountain, Canberra, ACT, Australia

Introduction

Experimental crosses have always been limited in size by the expense and time associated with phenotyping. For plants, this expense is incurred both through multiple experimental stages and the need for replicated designs at each stage. For example, study of baking quality will involve experimental stages such as the field, milling and baking; replication is critical to understand the sources of variation that contribute to the phenotype, such as spatial variation in the field. In contrast, since the advent of high-throughput genotyping, both time and cost required for genotyping have decreased greatly. Hence, a cost-effective strategy for phenotyping is to use available genetic information to select individuals with maximal potential.

A timely example of this phenomenon is embodied in the large complex crosses, which are in progress around the world. Examples include the Collaborative Cross in mice (The Complex Trait Consortium 2004), Nested Association Mapping populations in maize (Yu et al. 2008), and multi-parent advanced generation intercross (MAGIC) populations in plants (Cavanagh et al. 2008; Kover et al. 2009). Such populations are bred as a genetic resource for researchers to explore the underlying basis of many complex traits. Indeed, the final population is made up of inbred lines which need only be genotyped once. However,

individual studies will focus on specific traits and may not have the resources to phenotype the entire population. Hence, subsets of the population need to be selected for these studies based on genetic information, either from the whole genome or candidate regions.

The goal of identifying lines with maximum genetic diversity is also relevant to the selection of core collections in breeding populations, or in maximizing biodiversity in wild populations. Franco et al. (2006) suggest a general strategy to conserve diversity by clustering samples based on some distance measure and to then select representatives from each cluster as the subsample of interest. However, many distance measures which have been proposed (Mohammadi and Prasanna 2003; Reif et al. 2005) are inappropriate for experimental populations. For example, distance measures relying on population allele frequencies make no sense when the allele frequencies simply depend on the design. Similarly, measures of heterozygosity are inappropriate for dominant markers.

In experimental populations, the most basic form of selection uses no genetic information and selects individuals by simple random sampling (SRS) (Cochran 1977), where each individual has the same chance of being selected. However, Jin et al. (2004) demonstrated that their approach of minimum moment aberration (MMA) selects lines which have greater power to detect QTL in F2 populations compared with SRS. While this approach is not applicable to other designs and does not cope well with missing data, it nevertheless demonstrates the potential value of selective phenotyping. An alternative, maxRec, is to select the most recombinant progeny (Jannink 2005); however, this is limited to backcross, doubled haploid and bi-parental recombinant inbred lines (RIL). Ansari-Mahyari et al. (2009) compare random sampling with maxRec and other strategies using linkage and linkage disequilibrium characteristics to select phenotypes. In general, though, they find only moderate changes in power of QTL detection across strategies.

In this paper, we propose a general approach to selective phenotyping called SPCLUST. This approach accommodates practical issues in genetic studies such as missing data, dominant markers, and latent genotype data. It is applicable to general experimental designs, and we compare it to other selection methods through simulations of backcross, F2 intercross, and MAGIC 4-parent designs. For the MAGIC design, we extend the approach to encompass multiple stages of selection. Through simulation we show that selection within a previously detected QTL region has the potential to reduce the width of the QTL support interval and hence improve resolution for fine mapping. Finally, we apply SPCLUST to two examples of real datasets. The simulations and real data applications lead us to conclude that SPCLUST produces subsets of selected lines with higher diversity and similar or slightly higher power than other methods.

Materials and methods

We briefly review two other selective phenotyping methods which we will compare with the SPCLUST method proposed here. All three methods have similar aims; namely, to maximize the dissimilarity between selected individuals. However, the differences between the details of the approaches and the assumptions required may result in quite different selected samples upon application.

MMA (Jin et al. 2004)

This method minimizes the average of all pairwise similarities between individuals in the subsample, where similarities are based on the number of alleles shared at all markers. Forward selection of individuals, which are most dissimilar is followed by refinement of the sample through swapping of individuals. This approach was developed for F2 designs and has not been generalized to other experimental designs, nor can it be used with missing genotype data.

maxRec (Jannink 2005)

This method aims to maximize the overall mapping information available in the selected sample by maximizing the number of recombination events. Essentially, the number of intervals containing recombinations is counted for each individual, and the selected sample is made up of those individuals with the highest number of recombinations. This approach is limited to backcross, doubled haploid and bi-parental recombinant inbred lines (RIL), due to the difficulties in ascertaining recombination events in other experimental populations.

SPCLUST

We present here a method for selective phenotyping in general experimental designs. Given a population of size N, with each line genotyped at M markers, our goal is to select a sample of size n with maximum genetic diversity. Essentially, we aim to reduce the set of samples to those which are most different from each other so as to eliminate genetic redundancy. While this may not have maximum power for QTL detection since some genetic redundancy improves power, we will retain a set of lines for phenotyping which conserves the diversity of the whole population.

The basic idea of this approach is to

1. Compute genetic distances between all lines in a population.
2. Cluster based on the computed genetic distances to produce n clusters.
3. From each cluster, select a single representative.

All three steps can be performed in multiple ways. For example, in Step 3, we select the individual for which the sum of distances to all other members of the cluster is minimized, i.e. the most central member of the cluster. However, it would be equally possible to select one at random with equal probability. This alternative is effectively stratified random sampling where the strata are defined by the genetic information.

In addition to its use as a single-step selection process, SPCLUST can form the basis of multiple stages of selection. We denote this form of the algorithm as SPCLUST2. In particular, this usage is relevant to fine mapping: if a QTL is large enough to be detected in an initial genome-wide screen of n_1 lines, in a second stage of selection n_2 lines may be added based on genetic markers within the QTL region. Because the selected lines have higher genetic diversity than randomly selected lines, recombinations are more likely, and hence the resulting sample should provide better resolution of true location of the QTL.

For the second stage of selection, we again construct clusters of the relevant genetic distances, so that in total there are $n_1 + n_2$ clusters. Any lines selected in the original screen are by default included in the second stage; otherwise, the time and effort spent phenotyping the original sample would be wasted. From the clusters which contain no lines from the original screen, n_2 are randomly sampled; representative lines are then drawn from these clusters. Note that if the same genetic distance measure is used in both stages, then one representative will be drawn from each cluster; however, if different genetic distance measures are used, some clusters may not be represented.

This approach maximizes genetic diversity in the second sample, but we can also maximize the number of recombinations directly in a similar fashion to maxRec. For this, we first impute the most likely parental allele inherited at each marker and then estimate the number of recombination events for each line. We can use these estimates to directly select lines with the highest levels of recombination, ensuring as before that any lines selected in the original screen are included in our final sample.

Genetic distance measure

An integral part of the algorithm is the computation of genetic distance between lines. While in theory nearly any distance measure could be used, many are inappropriate for the conditions commonly encountered in experimental designs. For example, Euclidean distance, which computes the sum of squared differences between genotypes for two lines, cannot distinguish between "absence" values at a marker, although this category may encompass a variety of underlying genotypes. This issue can be overcome by using the Jaccard distance measure (Jaccard 1908), which considers only the "presence" values; however, by doing so we are neglecting part of the information in the data.

We define a general distance measure which is appropriate for various designs, missing data, dominant markers, and latent parental alleles. Let $\mathbf{X_i}$ and $\mathbf{X_j}$ denote the vectors of observed marker values for the i-th and j-th lines. We first compute the similarity coefficient between these two lines as the expected number of alleles in common between the two lines. The idea of computing the proportion of alleles shared over all loci has been proposed before (Bowcock et al. 1994); however, it is generally used in cases where there is no ambiguity about which alleles are inherited:

$$s_E = \frac{E\left[\mathrm{IBD}\left(\mathbf{X_i}, \mathbf{X_j}\right)\right]}{2M} = \frac{\sum_{k=1}^{M} \sum_{\mathrm{IBD}=y} yP\left(y \middle| X_{ik}, X_{jk}\right)}{2M}.$$

To calculate this expectation, we sum over all markers in the data and consider all identical-by-descent (IBD) possibilities for the individual marker values for lines i and j, multiplied by the probability of their occurrence. In practice, these probabilities are computed using the calc.genoprob function in R/qtl (Broman et al. 2003). For experimental crosses, this produces the probability of each potential genotype at a locus using hidden Markov model methodology.

For fully informative markers, where each founder exhibits a different allele, these probabilities are essentially 0 and 1. For backcrosses, the similarity coefficient reduces to the simple matching coefficient (Sneath and Snokal 1973), or the proportion of markers for which two lines have the same observed genotype. However, for dominant markers in an F2 intercross, these probabilities allow us to differentiate between the true underlying homozygous and heterozygous states. For biallelic markers in a MAGIC 4-way cross, we can distinguish between the probabilities that alleles are inherited from each of the four parents. Thus, knowledge of the structure of each breeding design provides additional genetic information to use in selection.

This similarity coefficient can be transformed to a distance measure $d_E = \sqrt{1 - s_E}$. After computing distances between all pairs of lines, we then input this distance matrix to clustering approaches.

Clustering

There are a variety of clustering methods which can be used with a given distance measure. Here, we consider two

hierarchical clustering approaches, as well as partitioning around medoids (PAM), a more robust version of k-means clustering (Kaufman and Rousseeuw 1990). The medoid is a logical representative from each cluster, analogous to the most central member of the cluster selected with hierarchical clustering.

For hierarchical clustering, we compare Ward's minimum variance method (Ward 1963) and the average linkage method (or unweighted paired group method using arithmetic averages—UPGMA) (Sneath and Snokal 1973). These two clustering approaches are the most commonly adopted agglomerative hierarchical methods and have been compared favourably to other approaches previously (Mohammadi and Prasanna 2003). We expect that different clustering approaches will result in different selected samples, since the distances between branches and general tree structures may be quite different.

Implementation

We have implemented SPCLUST in the R programming language (R Development Core Team 2011), making use of genetic imputation functions from R/qtl (Broman et al. 2003) and R/mpMap (Huang and George 2011). The package R/spclust, which includes both single and two-stage selection functions as well as visualization functions, is available on the Comprehensive R Archive Network (http://cran.r-project.org/web/packages/spclust/index.html).

Simulation

We demonstrate the robust nature of SPCLUST through simulation of different experimental designs. For each design type, we compare SRS, SPCLUST, and other methods which have been proposed for selection under that design. Further, for each design we consider a different facet of the data which could potentially affect the sampling results. In backcrosses, we consider the effect of missing genotypic data; in F2 intercrosses, we consider the effect of using dominant as opposed to codominant markers; and for the MAGIC 4-way design, we consider the effect of using biallelic as opposed to fully informative markers.

Data were generated using the packages R/qtl (Broman et al. 2003) for backcrosses and F2 designs and R/mpMap (Huang and George 2011) for the MAGIC design. The genetic map was simulated as containing five chromosomes of length 100 cM, with marker density varying between markers with equal spacing of 10, 4, and 2 cM. A QTL was generated on Chromosome 1 at 45 cM. For backcrosses and F2 designs, the QTL had effect size 1, explaining 20 and 33 % of the phenotypic variance, respectively. The full population contained $N = 200$ lines, from which $n = 100$ lines were selected in the backcross and $n = 50$ lines in the F2 population due to the larger proportion of variance explained by the QTL. For the MAGIC population, the QTL had effect size of 1.2, explaining 21 % of the phenotypic variance. Here, we have a larger population size of $N = 800$, but similar proportions of selected samples with sizes $n = 200$, 400, and 600. The increased population size for the MAGIC design reflects that populations are typically larger for this design. Each simulation scenario was repeated 500 times.

We compared methods in terms of genetic diversity of the selected lines and QTL mapping power. Genetic diversity was measured by the minimum distance between any pair of lines in the selected sample. For backcrosses and F2 designs, QTL mapping power was calculated as the percentage of replicates in which a QTL was detected within 10 cM to either side of its true location. For the MAGIC design, power is defined as the percentage of replicates where the QTL is detected within 5 cM of the true location. This more stringent requirement reflects the greater genetic resolution of the MAGIC 4-way design and its potential use for fine mapping. The threshold for QTL detection was determined as the 95th percentile of the maximum genomewide interval mapping test statistic. This was calculated by simulating 1,000 replicates generated from the null distribution of no QTL. Interval mapping was performed with the packages R/qtl and R/mpMap (Broman et al. 2003; Huang and George 2011) used to generate the data.

To test the efficacy of multiple stages of selection, we generate data as above for a MAGIC 4-way design with population size of 800 and marker density of 2 cM. A large QTL explaining 33 % of the phenotypic variance is generated at 45 cM. The first stage of selection proceeds exactly as before to select 100 lines. In the second stage, we perform selection via SPCLUST2 and by selecting lines with maximum recombinations (SPCLUST2-MR). Recombinations and the genetic distance between all lines in the population are computed using only genetic markers within 20 cM of the QTL detected in the first stage. An additional 100 lines are selected using the multi-stage approach.

Our comparison of these different methods is based on the width of the QTL support interval. We compute the QTL support interval as the region of the genome where the test statistic is within 12.6 of the maximum test statistic on the chromosome. Note that this corresponds roughly to the use of a 2-LOD support interval in standard QTL mapping; a 1-LOD support interval would be equivalent to a Wald test statistic within 7.81 of the maximum test statistic. While it is unclear what level of support corresponds exactly to a 95 % confidence interval, an appropriate value should fall somewhere in this range (Manichaikul et al. 2006).

We perform several comparisons with these data. First, we compare the two-stage selection approaches with a single-stage selection of 200 lines using SPCLUST. This allows us to determine whether the additional stage of selection does in fact result in improved mapping precision. Second, we compare these selection approaches with simple random sampling of 200 lines. Third, we generate SRS samples ranging in size from 200 to 800 in steps of 50. Comparing the selection approaches to this range of samples allows us to determine the effective sample size for each approach, i.e. what size random sample would be necessary to achieve similar results.

Data

Here, we consider the application of SPCLUST to two datasets. One is a population of $N = 176$ doubled haploid lines. These progeny were derived from the cross of the Australian bread wheat cultivar 'Chara' and the Canadian cultivar 'Glenlea'. Full details of the cross can be found in Cavanagh et al. (2010). A genetic map for this population was constructed in a previous study and contains 406 DArT, RFLP, and SSR markers (Huang et al. 2012a). The total map length was 4,658 cM resulting in an average density of one marker for every 11.5 cM. On average, 17.6, 28.3, and 12.1 markers were mapped to each chromosome in the A, B, and D genomes, respectively.

To illustrate the effect of selective phenotyping in such a population, we analysed data from a two-phase experimental design, with a field trial performed in Griffith, New South Wales, Australia and the 'b*' flour colour measured in the laboratory. This is an economically important trait describing flour yellowness which contributes to determining end use quality. Further description of the trial can be found in Huang and George (2009). We adopt a two-stage strategy to analyse the data. First, we construct a mixed linear model for the relationship between flour colour and environmental effects such as field spatial effects, design factors, and the milling process. Predicted means are computed from this model, which are then used as the response in the second stage of QTL mapping.

We compare results from the analysis of the full dataset with the results from three nested selected subsamples of size $n = 44$, 88, and 132 (25, 50, and 75 % of the total population size, respectively). For each sample size, lines are selected using SRS, MMA, and SPCLUST. The resulting datasets are analysed using composite interval mapping as implemented in R/qtl (Broman et al. 2003) with three cofactors.

The second dataset is from a four-way bread wheat MAGIC cross. Four Australian wheat cultivars (Chara, Baxter, Yitpi, and Westonia) were crossed in two pairs to create F1 seed Yitpi × Baxter (AB) and Chara × Westonia (CD). These F1 seeds were then intercrossed (AB × CD) via 70 independent crosses to generate 850 "4-way" (ABCD) F1 seed. Each ABCD F1 seed was grown, harvested and two F2 seeds progressed to the F6 generation by single seed descent to create $N = 1581$ recombinant inbred lines (RILs).

A field trial was conducted in 2009 at Yanco, New South Wales, Australia. The field trial consisted of 1,100 F6:8 RILs from the 4-way MAGIC population and check cultivars including the parental lines. The trial was based on a partially replicated (Smith et al. 2006) spatially optimized design using DiGGer (Coombes 2002). Forty percent of the RILs were replicated and check cultivars were sown in triplicate. In total, 1,620 plots were sown. Prior to machine harvest, plant height (cm) was recorded, and after harvest hectolitre weight was recorded (kg).

We analyse both traits in this dataset using mixed model interval mapping as implemented in R/mpMap (Huang and George 2011). In addition to genetic effects, all models contain fixed and random effects to account for the experimental design. At each potential QTL position, we test the joint (3 df) Wald statistic that all founder effects are zero. We consider datasets ranging in size from $n = 100$ up to $n = 1{,}000$ in step sizes of 100, as well as the full set of RILs. For each sample size, lines are selected using SRS and SPCLUST. For plant height, our aim is to determine how many lines are required to detect the QTL found in the full data. For hectolitre weight, our aim is to assess the support interval width for large QTL using different selection methodologies.

Results

Simulation

One aspect of the simulations was a comparison between different clustering methods used in SPCLUST—Ward, average, or PAM. In all situations, we found only small differences among these approaches and hence focus on average clustering in the results and figures. Extended results comparing all three approaches can be found in the Supplementary Material.

The first design considered is a backcross, or doubled haploid population, or biparental RILs; for our purposes, the only issue is whether the observed marker values exactly reflect the underlying genotypes. Generally, we see an improvement in both diversity (Fig. 1a) and power (Supp. Fig. 1) over random sampling for both SPCLUST and maxRec. However, the increase in power is not substantial for any method. Altering the marker density has only small effects. We found little difference between results when data were 10 % missing completely at random, when the missing data had been imputed, and when

the original full data were analysed. Indeed, we find that the selective phenotyping approaches tend to select lines with less missing data than average, which may explain the robustness to moderate levels of missing data.

The second design considered is an F2 intercross (Supp. Fig. 2). In this situation, MMA exhibits the highest power, while SPCLUST has similar power to SRS. Using dominant markers results in approximately 25 % loss in power for all methods. Although MMA produces samples with high power, it does not produce the most diverse samples. As was consistently true throughout simulations, SPCLUST produces samples which contain the most genetically diverse individuals (Fig. 1b). MMA-selected samples are similar in diversity to SRS for codominant markers and lower for dominant markers.

The third and final design considered is a MAGIC 4-way cross (Supp. Figs. 3, 4). No previously proposed methods are applicable here, so we compared SPCLUST only to random sampling. For this design, most genetic markers (including SNPs) will be incompletely informative due to the presence of more parents than alleles in the population. Figure 2 compares the performance of SPCLUST using fully informative markers to the more typical situation of biallelic markers for three different sampling proportions (25, 50, and 75 %). SPCLUST exhibits no real difference in power to SRS, but greater diversity. The decrease in diversity with sample size reflects the definition of diversity as the minimum distance between any members of the sample. Larger samples are thus more likely to contain similar individuals. There is an overall decrease in power for biallelic markers relative to fully informative markers. This is due to the greater uncertainty as to whether two individuals share alleles in the biallelic case. As the density of markers increases, the power approaches that seen for fully informative markers. This is to be expected since information is more closely shared between neighbouring markers.

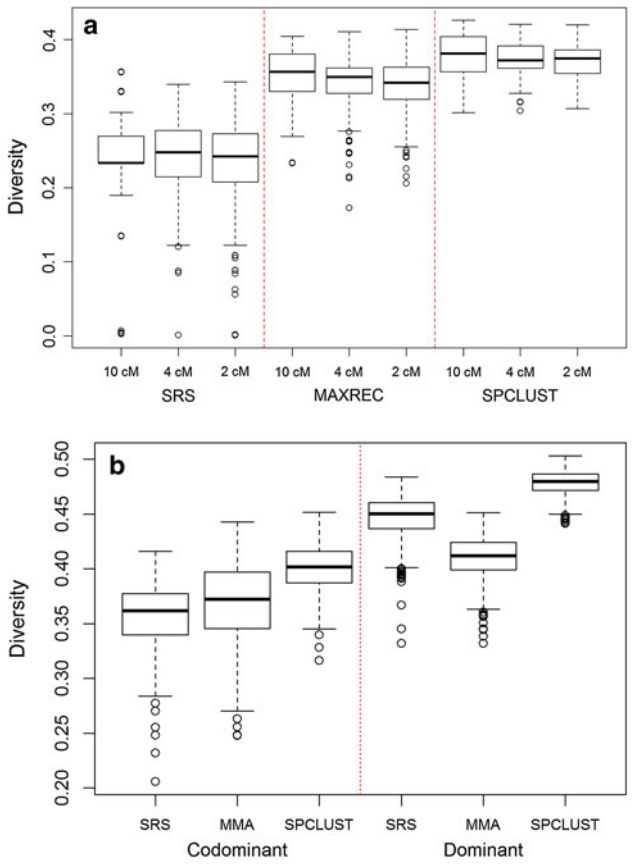

Fig. 1 Box plots of diversity scores in **a** selected subsets of size 100 from a 200-line backcross population at different marker densities and **b** selected subsets of size 50 from a 200-line F2 intercross population. In the F2 population, we compare results for codominant and dominant markers at a density of 4 cM. *Lines within boxes* indicate median values

We also explore the use of SPCLUST2 for fine mapping in MAGIC through multiple stages of selection. Because the QTL generated for this set of simulations explains a large proportion of the variance, it is detectable for over 90 % of replicates even with a sample size of 100. In replicates where the QTL is detected, we proceed to the second stage of selection to narrow the prospective QTL region. We compare the width of the QTL support intervals for SPCLUST2 and SPCLUST2MR using single and multi-stage selection against SRS for various sample sizes (Fig. 3). Both multi-stage procedures perform similarly, with a median width for SPCLUST2 of 5.16 cM and for SPCLUST2-MR of 4.74 cM. For SRS of the same size, the width tends to be over twice as large (11.8 cM), and even single-stage selective phenotyping performs much better with widths about 60 % as large. Indeed, the gain in resolution by using selective phenotyping is such that for a single stage, random samples of size 400 are required to reduce the support interval width comparably, while for multi-stage selection, the effective sample size is approximately 650.

The reason for this improvement becomes clear by considering the number of recombinations in the region surrounding the QTL (Fig. 4). Increased recombinations allow better resolution by increasing our ability to distinguish between the causal variant and nearby loci. Since SPCLUST2-MR directly selects lines with many recombinations, it is not surprising that it has the highest number of recombinations, averaging over 2/Morgan. SPCLUST2 is the next highest, averaging over 1.5/Morgan, while all other methods average about one. This increase in local recombination results from these being the only two methods, which select lines based on the genetic diversity in the region surrounding the QTL. One-stage SPCLUST does not particularly improve on SRS, since it does not focus on this specific region. However, its increased

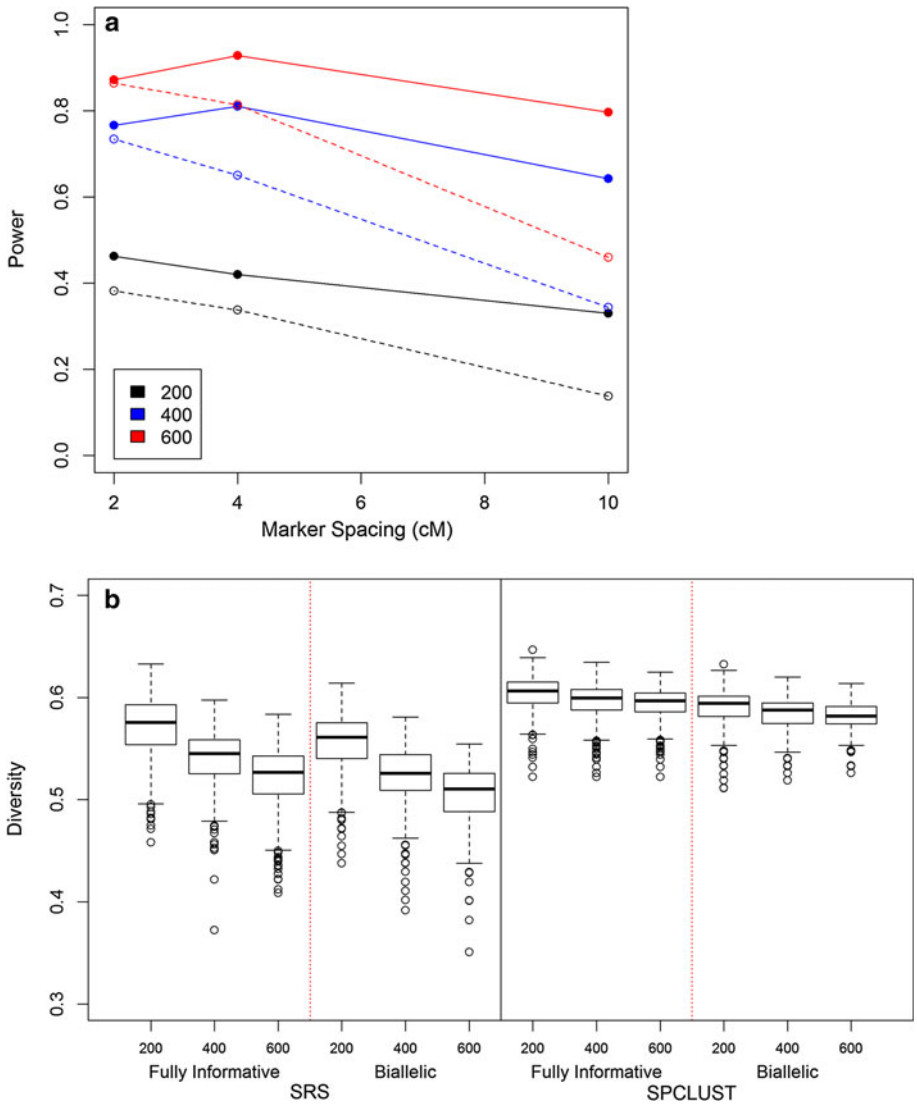

Fig. 2 For selected subsets of varying sample size (200, 400, 600) from an 800-line MAGIC 4-way population. **a** QTL mapping power at varying marker density. *Solid lines* denote the use of fully informative markers in selection and QTL mapping; *dashed lines* denote the use of biallelic markers. **b** Box plots of diversity scores (for marker density of 4 cM). *Lines within boxes* indicate median diversity scores

resolution for the QTL support interval is likely due to a general increase across the genome in recombinations due to the selection of lines with high genetic diversity.

Chara × Glenlea population

Previous composite interval mapping analysis of this trait on the Chara × Glenlea population (Huang and George 2009) detected QTL on chromosomes 4B, 7A, and 7B. Hence in our analysis, we are interested to see whether these QTL could be detected had selective phenotyping been used to reduce the number of individuals chosen from the population. Table 1 shows the maximum LOD score achieved on each of these chromosomes for different selection methods and sample sizes. In general, the LOD score increases with sample size. Using a significance threshold of 3 for the LOD score, all selection methods successfully identify the QTL on chromosome 7A. The QTL on chromosome 4B is only detected by SRS at a sample size of 88 individuals (50 % of population), while all methods detect it for the largest selected sample size of 132 individuals (75 % of population). The QTL on chromosome 7B is detected in all samples selected by SPCLUST, while the remaining methods detect this QTL as long as at least half the population is sampled.

MAGIC population

We analysed the traits 'plant height' and 'hectolitre weight' in the full dataset and selected samples of the MAGIC 4-way population. For plant height, at a significance threshold of 0.00017, we detected QTL on seven chromosomes in the full data: 1B, 2B, 2D, 4A, 4B, 4D, and 5B. The two largest QTL, on Chr 4B and 4D, represent known dwarfing genes (*Rht-B1* and *Rht-D1*) for the trait (Keyes et al. 1989), while the QTL on 2D may be related to the

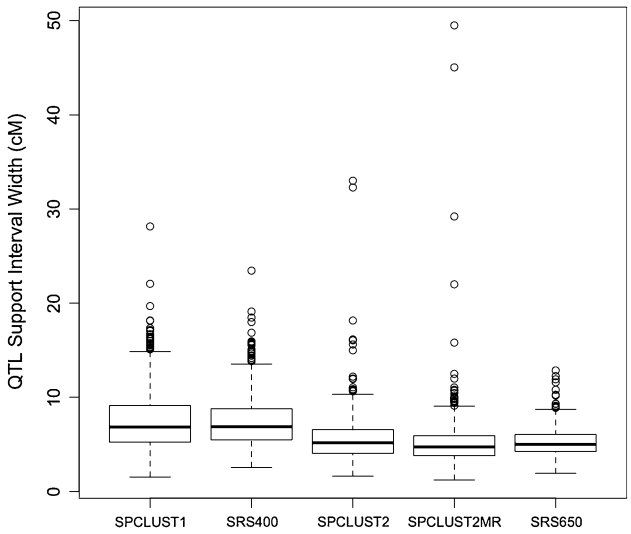

Fig. 3 Box plots of QTL support interval widths in samples of size 200 selected from an 800-line MAGIC 4-way population with a marker density of 2 cM. Selection was performed using single-stage SPCLUST1, multi-stage SPCLUST2, multi-stage SPCLUST2 with maximum recombination (SPCLUST2MR) and SRS. SRS samples of size 400 and 650 were closest in QTL support interval width to SPCLUST1 and SPCLUST2, respectively. *Lines within boxes* indicate median interval widths

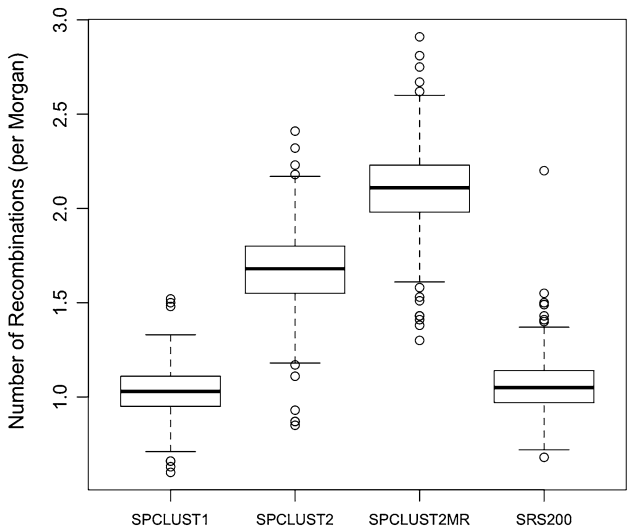

Fig. 4 Box plots of the number of recombinations in a region surrounding a QTL in samples of size 200 selected from an 800-line MAGIC 4-way population with a marker density of 2 cM. Selection was performed using single-stage SPCLUST1, multi-stage SPCLUST2, multi-stage SPCLUST2 with maximum recombination (SPCLUST2MR), and SRS. *Lines within boxes* indicate median number of recombinations

flowering gene for photo-period sensitivity *PPD-D1*. For hectolitre weight, we detected QTL on six chromosomes in the full data: 1A, 1B, 2B, 2D, 5A, and 7A (see Huang et al. (2012b) for full analyses). The largest QTL occurred on chromosome 2B in a region of markers associated with the alien introgression *Sr36* (Nyquist 1962). As this QTL is highly significant, it is detected in most smaller samples, so we can compare the precision of different methods based on its support interval width.

Table 1 Maximum LOD scores in interval mapping analysis of Chara × Glenlea population for different selective phenotyping methods

Chr	Sample Size	SRS	MAXREC	SPCLUST
4B	44	1.00	1.38	0.72
4B	88	**4.43**	2.21	2.18
4B	132	**4.13**	**3.13**	**3.06**
7A	44	**9.15**	**9.55**	**9.64**
7A	88	**19.18**	**16.39**	**17.21**
7A	132	**25.20**	**24.43**	**22.93**
7B	44	1.94	1.63	**4.44**
7B	88	**4.76**	**4.50**	**4.17**
7B	132	**5.15**	**5.31**	**4.99**

Values above an LOD threshold of 3 are in bold

For plant height, we selected samples ranging in size from 100 to 1,000 plants by steps of 100. In the sample of 100, SPCLUST detects a marginally significant QTL on chromosome 4D with p value 0.00037; SRS detects no QTL. In the sample of 200, SPCLUST detects significant QTL on Chromosomes 4B and 4D and a marginally significant QTL on Chromosome 2B ($p = 3.64e{-}4$); SRS detects only a QTL on Chromosome 4D. QTL profiles for the full analysis, SRS and SPCLUST are shown in Fig. 5. It is primarily in the smaller sample sizes that the benefits of selective phenotyping are realized, since for larger sample sizes, all sampling methods detect the QTL associated with the dwarfing genes, along with varying QTL on other chromosomes.

For hectolitre weight, we focus on samples of size 200 and 300 selected using single-stage and multi-stage SPCLUST as well as SRS. For multi-stage SPCLUST (SPCLUST2), we select additional lines based on genetic diversity in the 40 cM region surrounding the QTL detected in samples of size 100 on chromosome 2B. In a sample of size 200, SPCLUST2 detects the QTL on 2B in the correct region with a support interval width of 10 cM; SPCLUST detects the QTL with width 16 cM; and SRS detects the QTL with width 20 cM. For a sample of size 300, the width is 12 cM for SPCLUST2, 12 for SPCLUST, and 18 for SRS. For comparison, the support interval width in the full population is 10 cM.

Discussion

We have proposed a simple yet effective method of selective phenotyping, which is applicable and has robust performance for a wide variety of designs and data types.

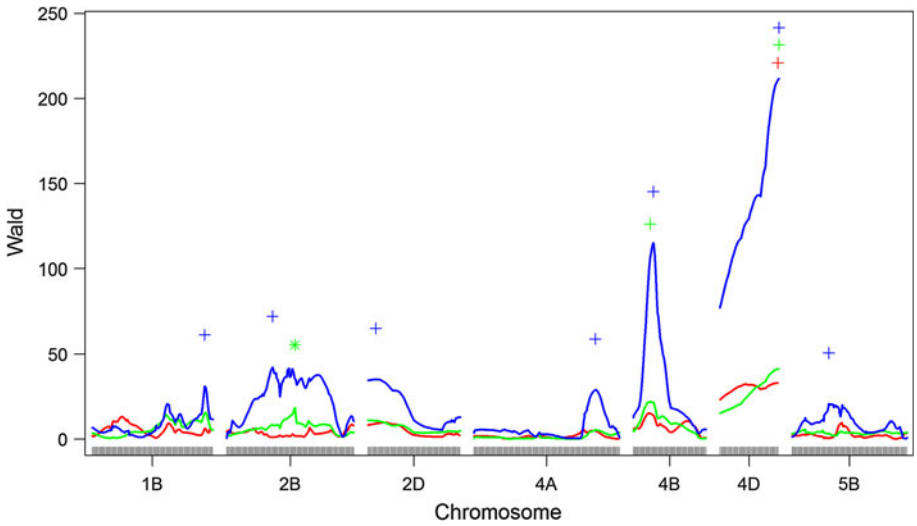

Fig. 5 Wald profiles for plant height (in cm) QTL analysis of selected chromosomes in subsample of size 200 of the MAGIC 4-way population. *Red lines* denote sample selected using SRS; *green lines* denote sample selected using SPCLUST; *blue lines* denote the full population. QTL detected by each method at p value thresholds of 0.001 and 0.00017 are indicated by *asterisks* and *pluses*, respectively (colour figure online)

The generalizability and flexibility of the SPCLUST approach make it a powerful tool in planning experimental cross studies. The approach lends itself naturally to answering questions about power and necessary sample sizes for phenotyping subsets of large populations. The robustness of the method makes it practical for all types of designs and genetic markers with no greater difficulty than random sampling.

SPCLUST has been developed to choose which lines to grow when a researcher has a specific sample size in mind. However, the approach can also be used to determine a suitable number of lines to select based on the genotypic information. An automatic method such as the gap statistic (Tibshirani et al., 2001) can estimate the optimal number of clusters based on a measure of within-cluster dissimilarity. Ideally, the number of lines grown should be greater than the number of natural clusters in the population. For example, if there are clearly 21 natural clusters in the population based on the genotypic information, then it would be foolish to miss out on a cluster because of an arbitrary cutoff of 20 lines. Of course, if the methods highlight 21 natural clusters and the researcher plans to run 30 lines, then the 9 additional lines are not to be considered a waste of resources; these additional lines will increase the overall power of detecting QTL.

We have compared SPCLUST with other sampling methods with regard to QTL mapping power and genetic diversity. SRS is a standard approach in agricultural studies, although various forms of selective phenotyping (Jin et al. 2004; Jannink 2005; Gagneur et al. 2011) have been proposed. Although these approaches all have high power in specific situations, none can be applied generally to experimental crosses. SPCLUST has none of the restrictions of other methods, but has comparable power and maximizes diversity amongst the chosen lines. In concept, it is most similar to Jannink's (2005) the maxRec method, which seeks to maximize the overall number of recombinations. This and SPCLUST are most useful when a genomewide analysis is undertaken, as the diversity across the whole genome is increased relative to random sampling, leading to improved resolution in QTL mapping. In contrast, MMA (Jin et al. 2004) is the most useful when selection is based on a few candidate genes. An intermediate measure between SRS and MMA might be to consider stratified random sampling based on candidate genes of interest. However, this requires prior knowledge about the trait of interest.

While we have proposed three different clustering methods for use with SPCLUST, in a real study a researcher can use only one. In general, we found that the average, Ward, and PAM clustering approaches performed similarly, and hence have focused on a single approach here and included more detailed comparisons in Supplementary Material. We did find that PAM had slight improvements over the other methods in terms of diversity and occasionally power, but we have not yet generalized this approach to multi-stage selection. However, we expect to observe the same improvements in resolution as for ward clustering in a multi-stage approach based on PAM. A further straightforward extension to the approaches which we have not considered here would be to incorporate information other than genotypes in the situation (e.g. a multi-stage trial) where some phenotypic information is available.

We have shown that multi-stage selection with SPCLUST is successful in increasing resolution and effective sample size in fine mapping. We expect this approach to be increasingly popular in large-scale studies aimed at narrowing down the QTL region to the gene level. This is an efficient way to exploit the relatively cheap genetic information available for large populations by reducing phenotyping requirements. While smaller samples will still

have reduced power relative to larger ones, they need not have decreased resolution in mapping detectable QTL. The drawback of this approach is the increased time required to complete the study; if phenotyping requires multiple stages of processing, such as for baking volume, then selection in two stages may be impractical. In that case, selection of the full sample using SPCLUST provides a compromise which balances time required against power and mapping resolution.

In addition to data examples, we have explored the characteristics of this selective phenotyping approach through simulation, as it is an inexpensive method of investigating a variety of different scenarios. The primary benefit of the simulation is that the underlying truth is known and we can control many factors to isolate their effect on different methods. We have focused on the effects of design type, marker type, and missing data here. In addition, during the simulation process we considered several different procedures for selecting a subset of markers to use in each method. In these cases, we found that varying our selection of markers or their density along a chromosome had little effect on the power and diversity of different methods and hence do not present the results.

References

Ansari-Mahyari S, Berg P, Lund MS (2009) Fine mapping quantitative trait loci under selective phenotyping strategies based on linkage and linkage disequilibrium criteria. J Anim Breed Genet 126:443–454

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890

Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11:215–221

Cavanagh CR, Taylor J, Larroque O, Coombes N, Verbyla AP, Nath Z et al (2010) Sponge and dough bread making: genetic and phenotypic relationships with wheat quality traits. Theor Appl Genet 121:815–828

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York. ISBN 0-471-16240-X

Coombes NE (2002) The reactive tabu search for efficient correlated experimental designs. Ph.D. thesis, Liverpool John Moores University, Liverpool

Franco J, Crossa J, Warburton ML, Taba S (2006) Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. Crop Sci 46:854–864

Gagneur J, Elze MC, Tresch A (2011) Selective phenotyping, entropy reduction, and the Mastermind game. BMC Bioinform 12:406

Huang BE, George AW (2009) Look before you leap: a new approach to mapping QTL. Theor Appl Genet 119:899–911

Huang BE, George AW (2011) R/mpMap: a computational platform for the genetic analysis of multi-parent recombinant inbred lines. Bioinformatics 27:727–729

Huang BE, Cavanagh C, Rampling L, Kilian A, George AW (2012a) iDArTs: increasing the value of genomic resources at no cost. Mol Breed. doi:10.1007/s11032-011-9676-5

Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR (2012b) A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol J. doi:10.1111/j.1467-7652.2012.00702.x

Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bul Soc Vaudoise Sci Natl 44:223–270

Jannink J-L (2005) Selective phenotyping to accurately map quantitative trait loci. Crop Sci 45:901–908

Jin C, Lan H, Attie AD, Churchill GA, Bulutuglo D, Yandell BS (2004) Selective phenotyping for increased efficiency in genetic mapping studies. Genetics 168:2285–2293

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Keyes GJ, Paolillo DJ, Sorrells ME (1989) The effects of dwarfing genes *Rht1* and *Rht2* on cellular dimensions and rate of leaf elongation in wheat. Ann Bot 64:683–690

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MC et al (2009) A multiparent advanced generation inter-cross to fine map quantitative traits in *Arabidopsis thaliana*. PLoS Genet 5:e1000551

Manichaikul A, Dupuis J, Sen S, Broman KW (2006) Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. Genetics 174:481–489

Mohammadi SA, Prasanna BM (2003) Analysis of genetic diversity in crop plants—salient statistical tools and considerations. Crop Sci 43:1235–1248

Nyquist WE (1962) Differential fertilization in the inheritance of stem rust resistance in hybrids involving a common wheat strain derived from *Triticum timopheevii*. Genetics 47:1109–1124

Reif JC, Melchinger AE, Frisch M (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Sci 45:1–7

R Development Core Team (2011) R: A language and environment for statistical computing R foundation for statistical computing, Vienna, URL http://www.R-project.org, ISBN 3-900051-07-0

Smith AB, Liw P, Cullis BR (2006) The design and analysis of multiphase plant breeding experiments. J Agric Sci 144:393–409

Sneath PHA, Snokal RR (1973) Numerical taxonomy. Freeman, San Francisco

The Complex Trait Consortium (2004) The collaborative cross, a community resource for the genetic analysis of complex traits. Nat Genet 36:1133–1137

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J Royal Statist Soc B 63:411–423

Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Statist Assoc 58:235–244

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551