ISSN 1479-2621

Plant Genetic Resources; 1–9
doi:10.1017/S1479262117000247

Comparison of optimization methods for core subset selection from a large collection of Mexican wheat landraces characterized by SNP markers

Carlos L. Acuña-Matamoros and M. Humberto Reyes-Valdés*

Departamento de Fitomejoramiento, Universidad Autónoma Agraria Antonio Narro, Buenavista, 25315, Saltillo, Coah., Mexico

Received 11 May 2017; Accepted 9 August 2017

Abstract

Core subset selection from collections hosted by seed banks, grow in importance as the number of accessions and genetic marker information rapidly increases. A data set of 20,526 single-nucleotide polymorphism (SNP) markers characterizing 7986 Mexican creole wheat landraces, was used to test 11 methods for core subset selection, through optimization criteria containing average genetic distance and genetic diversity. Allele richness was used as an additional criterion to qualify the generated core subsets. Three replications with random samples of 1500 SNP loci, each comprising a maximum of 3000 alleles, were used to perform the method evaluations through four different objective functions. The LR greedy search (LR) and LR with random first pair (LRSemi) were consistently best across all assays for maximizing the objective functions, and they performed well even for criteria not included in those functions. The Tukey's HSD (honest significant difference) multiple comparisons grouped those methods together with the sequential forward selection (SFS) and SFS with random first pair (SFSSemi) strategies as the top set of approaches. All of them are simple heuristic maximization algorithms, and outperformed two more sophisticated optimization approaches: parallel mixed replica exchange and replica exchange Monte Carlo. For their efficiency to optimize the objective functions and computing speed, the LRSemi and SFSSemi methods demonstrated to be good alternatives for core subset selection from large collections of highly homozygous accessions characterized by many biallelic markers.

Keywords: allele richness, diversity, genetic distance, seed banks

Introduction

Seed banks are aimed to maintain genetic diversity and to have ways for monitoring and evaluating populations (Govindaraj *et al.*, 2015). As seed banks grow in the quantities of accessions they handle, a necessity to select manageable subsets emerges. That motivated to define a core collection as a subset of accessions that represents, with minimum redundancy, the genetic diversity of the original collection (Frankel and Brown, 1984).

Several analytical strategies have been proposed for core subset selection, which use phenotypic, geographic and genotypic datasets. Those methods can be classified into two types. The stratified sampling strategy, which involves the conformation of groups by cluster analysis, and then sampling those clusters (Franco *et al.*, 1998), and direct optimization of some objective function applied to the whole collection, an approach that has demonstrated to be effective (Schoen and Brown, 1993; Thachuk *et al.*, 2009).

In the realm of direct optimization, Core Hunter (Thachuk *et al.*, 2009) was presented as a tool with an advanced stochastic local search (SLS) algorithm for selecting core subsets based on molecular markers. It aims to

*Corresponding author. E-mail: mathgenome@gmail.com

maximize an objective function that may contain several criteria. It uses an advanced SLS algorithm: replica exchange Montecarlo (REMC) (Geyer, 1991; Kimura and Taki, 1991; Iba, 2001). It performs replica subsetting, introduces perturbation and does replica exchanges with selection by the Metropolis criterion. This approach was tested with two molecular datasets of maize, comprising a maximum of 521 markers and 209 alleles, and it demonstrated to outperform MSTRAT (Gouesnard *et al.*, 2001), the D-method (Franco *et al.*, 2005) and Power Core (Kim *et al.*, 2007).

An innovation of the Core Hunter approach was presented with the Mixed Replica Algorithm (MixRep) (De Beukelaer *et al.*, 2012), which runs heterogeneous replicas. It was tested with plant datasets containing molecular marker data, with a maximum number of alleles of 282 and a maximum number of samples of 4429. It was shown that the MixRep gives cores with equal or higher diversity scores than REMC, and often outperforming it in computing speed. However, REMC and MixRep have not yet been tested with a large dataset of accessions and genetic markers, with a design suitable for hypothesis testing.

A large set of accessions characterized by many genetic markers is the collection of landraces introduced into Mexico from Europe. From them, 8616 have been characterized by Dart-seq technology, with availability of 20,526 quality single-nucleotide polymorphism (SNPs), as a part of the CIMMYT Seeds of Discovery initiative (Vikram *et al.*, 2016). Data are publicly available in the CIMMYT web page (http://www.cimmyt.org/) (Singh *et al.*, 2014). With this collection, core reference subsets have been assembled from the 7986 hexaploid accessions, by using a combination of SNP alleles, categorical and continuous phenotypic variables (Vikram *et al.*, 2016).

The objective of this work is to compare among eight simple heuristic optimization methods, plus REMC, MixRep and random sampling, for core subset selection from the large collection of Mexican wheat landraces characterized by SNP markers.

Materials and methods

Data from the collection of Mexican wheat landraces were downloaded from the web page http://www.cimmyt.org/ (Singh *et al.*, 2014). The subset of 7986 hexaploid accessions was selected, which correspond to the ones used by Vikram *et al.* (2016) in a study of genetic diversity of Creole wheats. Data are binary, with 0 denoting absence and 1 presence of a given allele. For homozygous loci, the binary notation was kept, whereas for heterozygous loci 0.5 was assigned to each allele, to be consistent with an allele frequency scale. For each direct optimization method tested and four objective functions (which we name 'conditions'), three replications were performed, each one based on a random sample of 1500 loci (3000 SNP alleles) extracted from the total of SNP 20,526 loci available in the dataset. Data filtering and random loci selection were performed with the aid of the language and environment for statistical computing R (R Core Team, 2016).

To perform optimizations for core subset selection, the software Core Hunter 2.0 (De Beukelaer *et al.*, 2012) was used through its R implementation, which has 13 search methods available. From them, two methods were excluded due to their high demand of time and system resources, which make them unsuitable for large datasets: exhaustive search and sequential backward selection. The following search strategies were tested: standard local search (Local), deterministic LR greedy search (LR), LR search with random first pair (LRSemi), parallel mixed replica search (MixRep), heuristic steepest descent (MSTRAT), random core set (Rand), Replica Exchange Monte Carlo (REMC), sequential forward selection (SFS), SFS with random first pair (SFSSemi), steepest descent search (Steepest) and tabu search (Tabu). For each assay, a core subset of 798 accessions was obtained, which represents nearly 10% of the collection.

For each tested method, four sets of weighting coefficients were used to include the following two criteria: modified Rogers Distance (MR) and Shannon diversity index (SH). The four objective functions were defined with 70% MR and 30% SH (default values in the software), 0% MR and 100% SH, 100% MR and 0% SH, and 50% MR and 50% SH. Since each of the four conditions was tested with three random marker sets, a total of 12 core subset selections were performed for each method. Four criteria were used to evaluate each core subset and the computer process: mean modified Rogers Distance between all pairs of accessions (MR), Shannon diversity index (SH), allele richness (AR) and computing time. Boxplots, multidimensional scaling graphics, hierarchical clustering, correlograms as well as Tukey's HSD (honest significant difference) tests were used to compare among the different methods.

The modified Roger's distance (Goodman and Stuber, 1983) is essentially Euclidean, and it is defined as follows:

$$\mathrm{MR} = \frac{1}{\sqrt{2m}} \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{k_i} (p_{ij} - q_{ij})^2},$$

where p_{ij} and q_{ij} are allele frequencies of the jth allele at the ith locus in two accessions under consideration, k_i is the number of alleles at the ith locus and m refers to the number of loci.

The so-called Shannon diversity index is originally the Shannon entropy (Shannon, 1948; Reyes-Valdés, 2013),

Table 1. Average scores for three core subsets generated by the use of ten methods plus random sampling, through optimization of objective functions containing a single criterion

Method	MR	SH	AR (%)	ObFun	ObFunSt	Time (s)
Objective = 0 MR + 1 SH						
Local	0.35233	7.63930	99.16	7.63930	0.82805	99
LR	**0.35240**	**7.63957**	**99.21**	**7.63957**	**0.84310**	17077
LRSemi	**0.35240**	**7.63957**	**99.21**	**7.63957**	**0.84310**	5989
MixRep	0.35170	7.63857	**99.21**	7.63857	0.78664	98
MSTRAT	0.32627	7.60570	98.15	7.60570	−1.06893	95
Rand	0.32477	7.60113	96.13	7.60113	−1.32676	31
REMC	0.33737	7.61987	98.76	7.61987	−0.26912	92
SFS	**0.35240**	**7.63957**	**99.21**	**7.63957**	**0.84310**	13249
SFSSemi	0.35230	7.63950	**99.21**	7.63950	0.83934	2999
Steepest	0.32513	7.60193	97.20	7.60193	−1.28159	3751
Tabu	0.32797	7.60627	98.13	7.60627	−1.03694	101
Collection	0.32491	7.60151	100.00	7.60151	–	–
Objective = 1 MR + 0 SH						
Local	0.33467	7.60933	97.18	0.33467	−0.44751	131
LR	**0.36173**	**7.63450**	**98.73**	**0.36173**	**1.12991**	8755
LRSemi	**0.36173**	**7.63450**	**98.73**	**0.36173**	**1.12991**	4856
MixRep	0.35427	7.62807	98.36	0.35427	0.69476	704
MSTRAT	0.32557	7.60133	97.41	0.32557	−0.97785	646
Rand	0.32440	7.60087	96.26	0.32440	−1.04584	69
REMC	0.32917	7.60410	96.65	0.32917	−0.76805	402
SFS	**0.36173**	**7.63450**	**98.73**	**0.36173**	**1.12991**	6622
SFSSemi	0.36167	7.63447	**98.73**	0.36167	1.12602	2814
Steepest	0.32537	7.60173	97.56	0.32537	−0.98951	4332
Tabu	0.32550	7.60147	97.12	0.32550	−0.98174	649
Collection	0.32491	7.60151	100.00	0.32491	–	–

The best score values are marked in bold.
MR, mean Modified Rogers distance; SH, Shannon diversity index; AR, Allele richness; ObFun, Objective function; ObFunSt, Standardized objective function.

and has the following expression for the ith locus:

$$\mathrm{SH}_i = -\sum_{j=1}^{k_i} p_{ij}\ln(p_{ij})$$

For optimization and evaluation, SH was calculated as the sum of SH_i across all loci for each sample of SNP markers. AR was calculated as the relative number of alleles in a core subset, compared with the number of alleles in the whole collection for the reference set of SNP markers. The values of AR are represented as percentages.

The optimization methods for core subset selection were run in a dual 2.5 GHz Intel Core i5 MacBook Pro processor, with 4 GB of RAM.

Results

Comparative scores

Results of the evaluations of the different search strategies through three random samples of 1500 SNP loci are summarized in Tables 1 and 2. When the objective function contained only SH (Table 1), the average value of this criterion was greatest for the LR, LRSemi and SFS methods, all of them being consistently best across the three replications (see online Supplementary material, Table S1). For AR, which was not included in the objective function, LR, LRSemi, SFS, SFSSemi and MixRep were the best performing, with each of them being consistently best in two of the three replications. For MR, which was not included in the objective function, the LR, LRSemi and SFS methods showed the highest average, and performed best in two

Table 2. Average scores for three core subsets generated by the use of ten methods plus random sampling, through optimization of objective functions containing two criteria

Method	MR	SH	AR (%)	ObFun	ObFunSt	Time (s)
Objective = 0.7 MR + 0.3 SH						
Local	0.33537	7.60997	96.95	2.51775	−0.42516	132
LR	**0.36133**	**7.63673**	**98.85**	**2.54395**	**1.11772**	21920
LRSemi	**0.36133**	**7.63673**	**98.85**	**2.54395**	**1.11772**	10973
MixRep	0.35387	7.63020	98.61	2.53677	0.69462	906
MSTRAT	0.32573	7.60123	97.07	2.50838	−0.97641	678
Rand	0.32580	7.60140	96.44	2.50848	−0.97072	73
REMC	0.32827	7.60347	96.60	2.51083	−0.83256	422
SFS	**0.36133**	**7.63673**	98.84	**2.54395**	**1.11772**	16630
SFSSemi	0.36130	7.63670	**98.85**	2.54392	1.11576	5744
Steepest	0.32570	7.60130	97.04	2.50838	−0.97660	8192
Tabu	0.32537	7.60177	97.12	2.50829	−0.98210	659
Collection	0.32491	7.60151	100.00	2.50789	–	–
Objective = 0.5 MR + 0.5 SH						
Local	0.33470	7.61050	97.50	3.97260	−0.42542	131
LR	**0.36053**	**7.63797**	**98.95**	**3.99925**	**1.11296**	19654
LRSemi	**0.36053**	**7.63797**	**98.95**	**3.99925**	**1.11296**	11622
MixRep	0.35327	7.63097	98.86	3.99212	0.70118	776
MSTRAT	0.32590	7.60160	97.25	3.96375	−0.93628	652
Rand	0.32477	7.59980	96.00	3.96228	−1.02095	68
REMC	0.32757	7.60400	97.32	3.96578	−0.81891	404
SFS	0.36050	**7.63797**	**98.95**	3.99923	1.11200	16734
SFSSemi	0.36050	7.63793	**98.95**	3.99922	1.11103	5385
Steepest	0.32523	7.60077	96.94	3.96300	−0.97958	8262
Tabu	0.32550	7.60087	97.17	3.96318	−0.96899	638
Collection	0.32491	7.60151	100.00	3.96321	–	–

The best score values are marked in bold.
MR, mean Modified Rogers distance; SH, Shannon diversity index; AR, allele richness; ObFun, objective function; ObFunSt, standardized objective function.

of the three replications, with the Local method giving the best score for Replication 2. When the objective function contained only MR (Table 1), the average value of this criterion was the highest for the LR, LRSemi and SFS methods, while they were matched by SFSSemi in Replication 1. For AR, the LR, LRSemi, SFS and SFSSemi methods showed the highest averages. For SH, which was not included in the objective function, LR, LRSemi and SFS showed the best averages, with the three of them being consistent across the three replications.

In the software default option, which assigns weights of 70 and 30% to MR and SH, respectively (Table 2), the LR, LRSemi and SFS methods showed the best average scores for the value of the objective function. For Replication 3, those methods were matched by SFSSemi. For AR, the LR, LRSemi and SFSSemi methods outperformed all other strategies, with the three of them being consistent across the three replications. When MR and SH were equally weighted (Table 2), the LR and LRSemi methods showed the best averages for the objective function, being matched by SFS and SFSSemi in two replications. For AR, the LR, LRSemi, SFS and SFSSemi methods had the highest average scores, although MixRep performed best for AR in the third replication.

The LR and LRSemi methods had the best scores for the objective function in the 12 assays, while they were matched by SFS in 11 of 12 assays. Although AR was not part of the objective functions, LR and LRSemi performed best for this criterion in 10 of 12 assays. In general, the best performing methods for the criteria contained in the objective functions were LR, LRSemi and SFS, with the fastest one of them being LRSemi. However, SFSSemi, which showed high scores too, is approximately twice as fast as LRSemi.

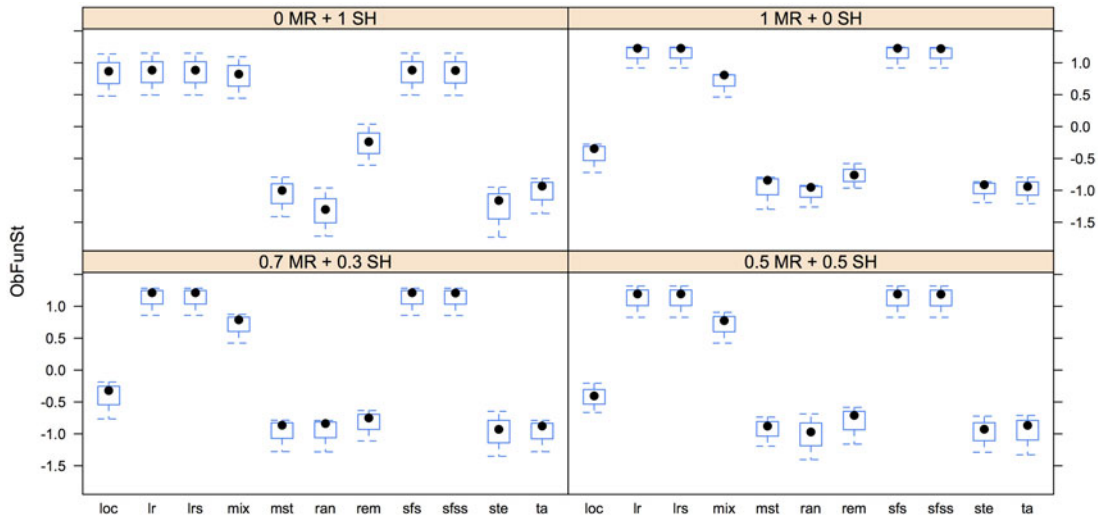

Fig. 1. Boxplots of the standardized objective function reached by different methods for core subset selection within each optimization condition, based on three replications. The methods are coded as follows: loc, Local; lr, LR; lrs, LRSemi; mix, MixRep; mst, MSTRAT; ran, Rand; rem, REMC; sfs, SFS; sfss, SFSSemi; ste, Steepest and ta, Tabu.

Table 3. Results of Tukey's HSD multiple comparisons for the performance of ten optimization methods for core subset selection plus random sampling

Method	ObFun	SqrObFun	Group	Time
LR	3.63613	1.74005	a	16851
LRSemi	3.63613	1.74005	a	8360
SFS	3.63612	1.74005	a	13309
SFSSemi	3.63608	1.74003	a	4236
MixRep	3.63043	1.73744	b	621
Local	3.61608	1.73058	c	123
REMC	3.60641	1.72753	d	330
MSTRAT	3.60085	1.72578	d	518
Tabu	3.60081	1.72575	d	512
Steepest	3.59967	1.72552	d	6134
Rand	3.59907	1.72523	d	60

ObFun, objective function; SqrObFun, square root of objective function.

The MR and SH values of the whole collection, reported in Tables 1 and 2, cannot be considered as the theoretical maxima for core subsets, because they do not necessarily grow with the number of accessions. In fact, they are surpassed by most of the core subsets. The only parameter that reach its maximum in the whole set of accessions is AR, with 100%.

To investigate the origin of collateral effects of optimization, where criteria appeared relatively maximized even when they were not part of the objective function, the 132 core subsets generated in this research were used to estimate correlations among the three criteria employed here. The three correlations were statistically significant ($P < 2.2 \times 10^{-16}$): MR with SH, $r = 0.96$; MR with AR, $r = 0.82$; and SH with AR, $r = 0.87$. These results explain the collateral effects of the different objective functions defined for the set of comparisons.

Boxplots for the standardized value of the objective function reached by each method within each condition, are depicted in Fig. 1. A similar pattern can be observed for the conditions 1 MR + 0 SH, 0.7 MR + 0.3 SH, and 0.5 MR + 0.5 SH, with the LR, LRSemi, SFS and SFSSemi methods standing on top of the plot, MixRep appearing close below them, the Local method showing up at the middle, and the remaining methods appearing as a group at the bottom of the plot. However, this pattern is different for the condition 0 MR + 1 SH, i.e. when only the Shannon diversity was used as a maximization criterion. In this situation, Local, LR, LRSemi, MixRep, SFS and SFSSemi appear at the top, REMC appears at the middle and the remaining methods at the bottom.

The analysis of variance for a model containing the terms method, condition and method × condition gave highly significant results for the three sources of variation, always with $P < 2.2 \times 10^{-16}$. For this analysis, the response variable, i.e. the objective function, was transformed to its square root, to approach a normal distribution for the ANOVA residuals. In Table 3, the average values of the objective function and they square roots, are shown for each method, along with the grouping for a Tukey's HSD test performed with an alpha value of 0.05. The group composed by the LR, LRSemi, SFS and SFSSemi methods showed the maximum efficiency, with LR and LRSemi having identical results for the average objective function. In second place appears the MixRep method, showing a

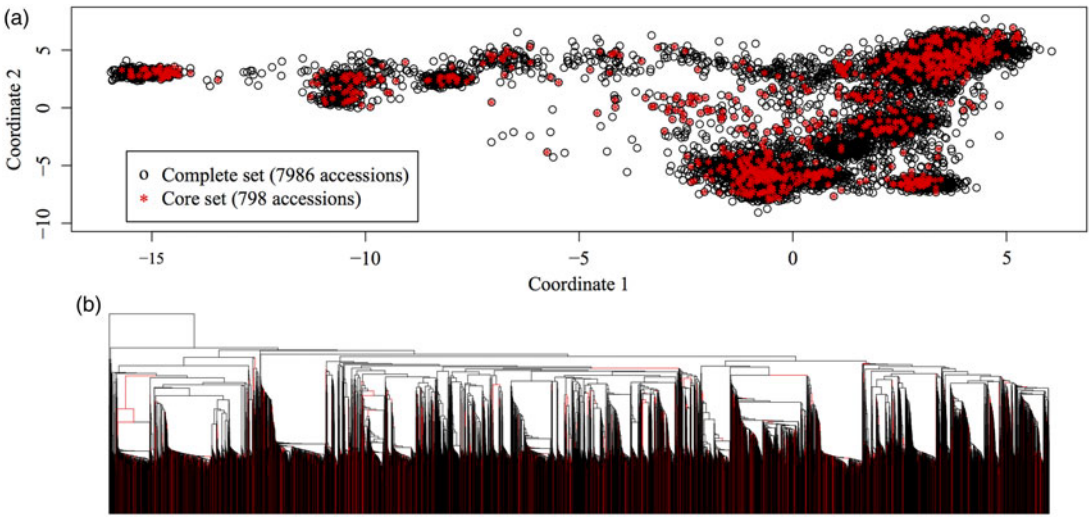

Fig. 2. Representation of a core subset composed by 798 accessions, in the context of the 7986 lines that integrate the whole collection. The optimization was carried out through the LRSemi method, based on the information of 1500 SNP loci. (a) Multidimensional scaling, with the original accessions being represented by black circles, and the selected accessions by red asterisks. (b) Hierarchical UPGMA cluster with the original accessions being represented by black lines, and the selected accessions by red lines.

behavior above the average of all strategies. In third place appears the Local method, with a behavior slightly less than average. Finally, in fourth place appears the group composed by REMC, MSTRAT, Tabu, Steepest and Rand. As expected, the Rand method had the lowest observed average score. Among the members of the first statistical group, the fastest method was SFSSemi. However, between the two methods that always performed best, the fastest was LRSemi. The MixRep method, which conforms the second statistical group, was faster than all members of the first group; in fact, it was 13.5 times faster than LRSemi. For the remaining two groups, the Local strategy was the fastest, after discarding the Rand method, which is not an optimization *per se*.

Attributes and coincidences of the core subsets

Figure 2 represents a core subset generated by the LRSemi strategy, which had the best scores along with LR in all assays, but with a higher speed. For this representation, the core subset was generated with an objective function that assigns the same weights to MR and SH. The multidimensional scale plot generated with 1500 SNP loci in Fig. 2 (a), depicts all the accessions in the collection, with the members of the core subset in red asterisks. Although this is a bi-dimensional representation of a highly dimensional space, it shows that the members of the core subset represent well the space of genetic variation. A different approach for graphic representation is depicted in Fig. 2(b), through an UPGMA hierarchical cluster plot. Although a large amount of wheat lines is represented in this plot, it can be noted that all groups tend to be represented in the core subset.

To evaluate the coincidence between the core subsets generated by the various strategies evaluated through this research, the number of common accessions was quantified for the subsets selected by the different methods in Replication 1. For instance, when the objective function contained both criteria, MR and SH, equally weighted (see online Supplementary material, Table S2), the LR and LRSemi methods had a coincidence of 100% between them, and composed a group with SFS and SFSSemi with coincidences of 99.62% and above. The MixRep strategy generated a core set with coincidences above 55% with the best four methods. The Local strategy had coincidences above 16% with the members of the group composed by LR, LRSemi, SFS, SFSSemi and MixRep. The group that showed the lowest scores, composed by REMC, Steepest, MSTRAT, Tabu and Rand, had a maximum coincidence of 11.4% with all methods. Correlograms that represent the coincidences among all methods for Replication 1, under all conditions, are depicted in Fig. 3. A consistent pattern emerged for all conditions, although MixRep was remarkably coincident (>80%) with LR, LRSemi, SFS and SFSSemi, when SH was the only component of the objective function (Fig. 3(a)), an attribute consistent with the boxplot in the left upper corner of Fig. 1. The general grouping that emerges from core subset coincidences resembles the structure given by the Tukey's HSD test.

A core set with 1133 lines was generated by LRSemi with the first sample of 1500 SNP loci, assigning equal weights to

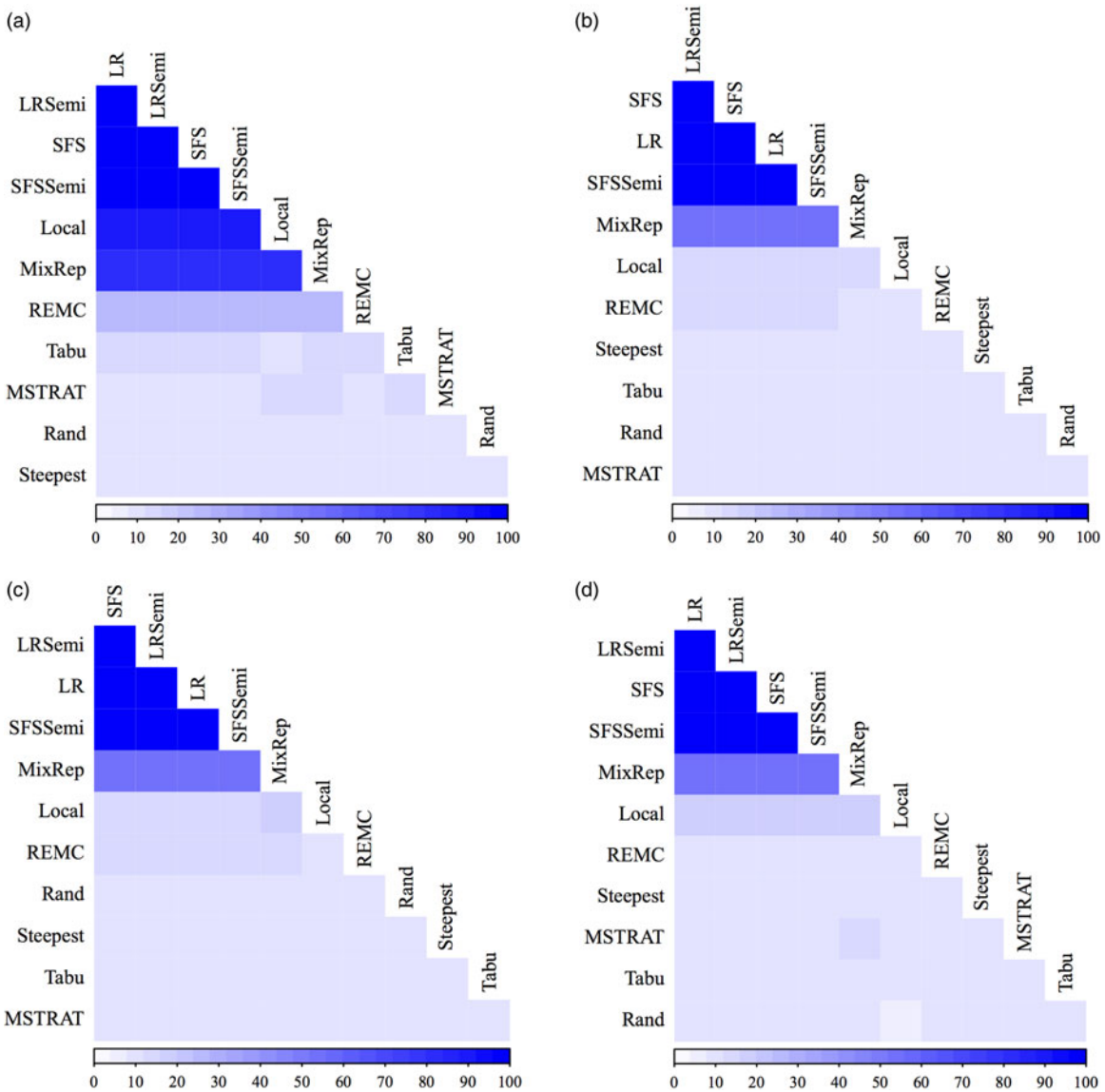

Fig. 3. Correlograms showing the percentage of coincidence among core subsets generated by the evaluated methods. (a) Objective = 0 MR + 1 SH, (b) Objective = 1 MR + 0 SH, (c) Objective = 0.7 MR + 0.3 SH, (d) Objective = 0.5 MR + 0.5 SH.

MR and SH, to compare it with the one reported by Vikram *et al.* (2016). A total of 171 out of 1133 accessions were common between both cores, i.e. a coincidence of 15%. Although this value is low, we must consider that in the core subset generated in the cited work, they used a combination of SNP markers and categorical and continuous phenotypic traits. Furthermore, in that paper the SNP information was not used directly, but reduced to 2000 principal components, and then six principal axes of a hierarchical multiple-factor analysis (HMFA) were selected to represent genotypic and phenotypic variance in proportions of 75 and 25%, respectively. They followed a stratified sampling strategy with the number of accessions per group estimated by the D-method. The scores of the published core for the reference set of 1500 SNP loci were MR = 0.322, SH = 7.595 and AR = 97.59%. For the core generated by LRSemi, the scores were MR = 0.353, SH = 7.627 and AR = 99.09%. Although the core subset selected by LRSemi showed better MR, SH and AR values, we must take into account that both cores were selected under different criteria.

Discussion

The LR and LRSemi methods performed consistently well in all assays, and they worked well even for the criteria not included in the objective function, due to correlations among the respective parameters. They performed identically, but

with LRSemi being considerably faster. Close to them, forming a single statistical group, appear SFS and SFSSemi as efficient algorithms to maximize the objective function, with SFSSemi being the faster. The MixRep method, although in the second statistical group, could be considered as a good choice only when computer time is an important limitation, also it is among the best strategies when the objective function is formed only by SH. When time is not extremely critical, LRSemi can be considered as a standard for the criteria evaluated in this work, because it consistently showed a good performance and moderate speed. The Local strategy, although a very fast method, does not appear to be a good maximization choice, because it had a less than average efficiency scores. Finally, at least under the characteristics of the data herein analysed, the REMC, MSTRAT, Tabu and Steepest strategies, do not seem to be a good choice, because they form a single statistical group with the use of a random extraction. It is worth remarking that the best performing methods collaterally aided to maximize criteria not included on the objective function.

The LR search, which was always consistently at the top of efficiency, is a greedy deterministic algorithm that does not take any random decision. It starts with the empty solution and iterates according to pre-determined parameters l and r, until the desired core size has been reached. In the implementation for Core Hunter 2.0, the conditions of the search are set in $l=2$ and $r=1$. In the LRSemi strategy, which performed as good as LR, the first pair of accessions is chosen randomly, thus being a semi-deterministic algorithm. As we have seen in the Results section, these two methods performed always on top of efficiency for the objective function. In spite of the high dimensionality of the genetic marker data, the graphic analysis offers an indication of a good coverage of the core set selected by LRSemi in the space of all accessions. Even when LRSemi is not totally deterministic, it showed all times a 100% coincidence with LR. On the other hand, the structure of the coincidence matrices for all methods reinforces the statistical grouping found by the Tukey's HSD test, being a good indication of the consistency of our results. Then, we must conclude that the group formed by the strategies LR, LRSemi, SFS and SFSSemi is a set of effective and consistent methods for selecting core subsets, at least from large collections of highly homozygous lines, genotyped with a high coverage by biallelic markers, as is the case of SNPs.

Small variations in MR, SH and AR are typical when comparing core selection methods (see for example De Beukelaer *et al.*, 2012); however, our design allowed discrimination among several strategies, even when the size of the core was large. Also, one must consider that small variations, e.g. in AR, can have a considerable impact on the richness of the subset. For example, a difference in 3% for AR between two methods can involve a difference up to 90 alleles between the two respective cores for the number of SNPs considered in this work.

To the best of our knowledge, this comparison of optimization approaches for core subset selection with genetic marker data is so far the one performed with the largest dataset, both in number of accessions and marker alleles. In fact, each of the three assays used herein, were performed on data sets with 23,958,000 points, while the largest reported in De Beukelaer *et al.* (2012) contained 200,364, from the flax data set comprising 708 samples and 282 alleles. In that work, it was observed that REMC was outperformed by simpler methods in several experiments. In fact, REMC never outperformed LR, which is consistent with our results. For the larger pea dataset REMC and MSTRAT resulted in worse scores than Local search. They also found the Local search to be faster than the REMC method. When Local, MSTRAT, LR and REMC were compared for minimum distance, LR showed leading results too. Although those results are coincident with our findings, the comparisons between Local, MSTRAT, LR and REMC were not as contrasting in De Beukelaer *et al.* (2012) as in this work, probably because of the much larger data set we used. In addition, in this work MixRep was for the first time systematically compared with the simple heuristics. Furthermore, our replication schema allowed us to perform statistical comparisons between methods.

The core subset of 1133 accessions that we selected by the LRSemi strategy, to make a comparison with the one selected by Vikram *et al.* (2016), showed the efficiency of LRSemi, albeit the selection criteria were different.

One of the justifications for the development of the MixRep method, pointed out by in De Beukelaer *et al.* (2012), is that LR becomes slow when run for relatively large datasets. However, as shown by our results, LRSemi performed as good as LR in large datasets, and took an average of 139 min to select core subsets with a selection pressure of 10% from tables of 7986 accessions with 1500 SNP loci, being a reasonable amount of computing time. As is the case of MixRep, the LR and LRSemi methods can take objective functions with other criteria, like minimum distance.

The increasing computing power is making feasible to efficiently apply direct heuristic algorithms for core subset selection, which can be efficient even for large data sets, and that can outperform complex strategies aimed to expedite the computing process. The LRSemi and SFSSemi methods are examples of simple heuristic algorithms that can be effectively used to develop core subsets from large collections, characterized by many marker loci.

Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1017/S1479262117000247.

Acknowledgements

This research was funded by Universidad Autónoma Agraria Antonio Narro. We thank Consejo Nacional de Ciencia y Teconología of Mexico for granting a graduate scholarship to C.L. Acuña-Matamoros, and the MasAgro and CIMMYT Seeds of Discovery initiative for allowing access to the wheat SNP data.

References

De Beukelaer HD, Smýkal P, Davenport GF and Fack V (2012) Core Hunter II: fast core subset selection based on multiple genetic diversity measures using Mixed Replica search. *BMC Bioinformatics* 13: 312.

Franco J, Crossa J, Villaseñor J, Taba S and Eberhart SA (1998) Classifying genetic resources by categorical and continuous variables. *Crop Science* 38: 1688–1696.

Franco J, Crossa J, Taba S and Shands H (2005) A sampling strategy for conserving genetic diversity when forming core subsets using genetic markers. *Crop Science* 46: 854–864.

Frankel OH and Brown AHD (1984) Plant genetic resources today: a critical appraisal. In Holden JHW and Williams JT (eds) *Crop Genetic Resources: Conservation and Evaluation*. London: George Allen and Unwin, pp. 249–257.

Geyer CJ (1991) Markov chain Monte Carlo maximum likelihood. In Keramidas (ed.) *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Interface Foundation: Fairfax Station, pp. 156–163.

Goodman MM and Stuber CW (1983) Races of maize: vI. Isozyme variation among races of maize in Bolivia. *Maydica* 28: 169–187.

Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ and David JL (2001) MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. *The Journal of Heredity* 92: 93–94.

Govindaraj M, Vetriventhan M and Srinivasan M (2015) Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genetics Research International* 2015: 14.

Iba Y (2001) Extended ensemble monte carlo. *International Journal of Modern Physics C* 12: 623–656.

Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, Gwag JG, Kim TS, Cho EG and Pak YJ (2007) Powercore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* 23: 2155–2162.

Kimura K and Taki K (1991) Time-homogeneous parallel annealing algorithm. In Vichneetsky R and Miller JJH (eds.) *Proceedings of the 13th IMACS World Congress on Computation and Applied Mathematics (IMACS'91)*, vol. 2. Dublin, Ireland: International Association for Mathematics and Computer Simulation, pp. 827–828.

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/ (Accessed January 2016).

Reyes-Valdes MH (2013) Informativeness of microsatellite markers. In: Kantartzi SK (ed.) *Microsatellites. Methods in molecular biology (Methods and Protocols)*, vol. 1006. Totowa NJ, USA: Humana Press, pp. 257–270.

Schoen DJ and Brown AHD (1993) Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. *Proceedings of the National Academy of Sciences of the United States of America* 90: 10623–10627.

Shannon CE (1948) A mathematical theory of communication. *The Bell System Technical Journal* 27: 623–656.

Singh S, Sansaloni C, Petroli C, Ellis M and Kilian A (2014) DArTseq-derived SNPs for wheat Mexican landrace accessions International Maize and Wheat Improvement Center (CIMMYT). Available at http://hdl.handle.net/11529/10013 (Accessed September 2015).

Thachuk C, Crossa J, Franco J, Dreisigacker S, Warburton M and Davenport GF (2009) Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. *BMC Bioinformatics* 10: 243.

Vikram P, Franco J, Burgueño-Ferreira J, Li H, Sehgal D, Saint Pierre C, Ortiz C, Sneller C, Tattaris M, Guzman C, Sansaloni CP, Ellis M, Fuentes-Davila G, Reynolds M, Sonder K, Singh P, Payne T, Wenzl P, Sharma A, Bains NS, Singh GP, Crossa J and Singh S (2016) Unlocking the genetic diversity of Creole wheats. *Scientific Reports* 6: 23092.