Scientific Research Publishing

Open Journal of Statistics, 2023, 13, 746-760
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

An Adaptive Sequential Replacement Method for Variable Selection in Linear Regression Analysis

Jixiang Wu*, Johnie N. Jenkins, Jack C. McCarty Jr.

Genetics and Sustainable Agriculture Research Unit, USDA-ARS, Mississippi State, USA
Email: *Jixiang.wu@usda.gov

How to cite this paper: Wu, J.X., Jenkins, J.N. and McCarty Jr., J.C. (2023) An Adaptive Sequential Replacement Method for Variable Selection in Linear Regression Analysis. *Open Journal of Statistics*, **13**, 746-760.
https://doi.org/10.4236/ojs.2023.135036

Received: September 25, 2023
Accepted: October 22, 2023
Published: October 25, 2023

Abstract

With the rapid development of DNA technologies, high throughput genomic data have become a powerful leverage to locate desirable genetic loci associated with traits of importance in various crop species. However, current genetic association mapping analyses are focused on identifying individual QTLs. This study aimed to identify a set of QTLs or genetic markers, which can capture genetic variability for marker-assisted selection. Selecting a set with k loci that can maximize genetic variation out of high throughput genomic data is a challenging issue. In this study, we proposed an adaptive sequential replacement (ASR) method, which is considered a variant of the sequential replacement (SR) method. Through Monte Carlo simulation and comparing with four other selection methods: exhaustive, SR method, forward, and backward methods we found that the ASR method sustains consistent and repeatable results comparable to the exhaustive method with much reduced computational intensity.

Keywords

Adaptive Sequential Replacement, Association Mapping, Exhaustive Method, Global Optimal Solution, Sequential Replacement, Variable Selection

1. Introduction

With the rapid development of DNA technologies, high throughput genomic data has been becoming a powerful leverage to locate desirable genetic loci associated with traits of importance in various crop species. It is well known that many quantitative traits like crop yield, plant height, and seed quality are controlled by many individual quantitative trait loci (QTLs) with minor effects and

possible interactions with environmental conditions. Current genetic association mapping is focused on identifying individual QTLs. Therefore, it is crucial to identify a set of QTLs, which can capture sufficient genetic variability for marker-assisted selection. Selecting a set of loci that can maximize genetic variation out of high throughput genomic data is desired but still computationally challenging.

Simple interval or composite interval mapping were commonly used to identify QTLs for controlled mapping populations (like F2, RI, or DH) when linkage maps are available [1]-[7]. These methods aim to identify each individual QTLs with integrations of linear regression and expected maximum (EM) algorithm when a linkage map is available [4] [7]. When a linkage map is constructed from high throughput genomic data, the distance of two flanking marker loci is often less than 2 centimorgan (cM), a window size commonly used by interval mapping may not be required. Genome-wide association studies (GWAS), on the other hand, have been focused on identifying individual genetic loci attributing phenotypic variation for an uncontrolled/random mapping population with and without population structure [8] [9].

Due to the potential of linked or interactive QTLs, the total amount of heritability of a set of QTLs is sometimes not the cumulation of the heritability of each identified individual QTLs. Therefore, it is important to select a set of loci that can catch the maximum genetic variation for a trait of interest. Such a process becomes the variable selection process in multiple linear regression, which aims to select the best subset of k variables out of the total p candidate independent variables. Given p genetic markers/loci, there are $(2^p - 1)$ all possible linear models to be examined. There is no doubt that the all-possible regression approach (sometimes called exhaustive method, which would be used throughout this study for consistency) is best because it examines every possible model [10]. However, a serious challenge associated with the exhaustive method is that the number of all-possible models could be very large even for a small number of independent variables [11]. Because of the high computational demand associated with the exhaustive method, heuristic methods are more frequently used for variable selection in linear regression analysis. They include forward selection (FS), backward elimination (BE), and stepwise selection (SS) [12] [13], which are currently available in several popularly used computer tools in R like MASS, leaps, and olsrr [14] [15] [16]. Although these variable selection procedures are very popular in the literature, a considerable number of limitations have also been identified due to the collinearity and/or interactions among predictable variables [17] [18] [19] [20]. For example, an excellent model could be overlooked by these selection methods because of the restriction of adding/deleting only one variable at each time and thus these procedures may not always yield the optimal regression model [18] [19]. Such findings were reported in Chapter 3 in the book of "Subset Selection in Regression Analysis, 2nd Edition" (Miller, 2002).

The number of variables being significantly selected by many variable selec-

tion methods could be large. Mathematically, it is desired to predict a response variable using more significant contributing variables. Sometimes, however, a plant breeder may be interested in identifying only four or five genetic markers rather than all contributing markers for a marker-assisted selection (MAS) practice. Therefore, selecting k variables (where k is given like 4 or 5), which aims to seek the smallest residual sum of squares (RSS) or the largest coefficient of determination (R^2), could be another desirable option for a breeding practice. This procedure will require a total of C_p^k equations to be examined to identify the best k-variable model if exhaustive method is applied. For example, a global search of the best subset of five ($k = 5$) variables out of 100 ($p = 100$) will need to examine over 75 million models.

In order to avoid the high computational demand associated with the exhaustive method, many scientists developed other alternative variable selection methods to improve the possibility to search the best subset for a given size of k variables [21] [22] [23] [24] [25]. Among these, a sequential replacement (SR) algorithm was proposed to improve variable selection with much reduced computational intensity [11] [18] [26] [27] [28] and the SR method is available in a popularly used R package (leaps) [14]; however, we discovered that the SR method could sometimes yield some inconsistent or undesired results, as demonstrated in this study. Therefore, it is important to improve the SR method so that both the power and speed can be sustained to achieve an optimal k-variable model selection.

In this study, our first objective was to propose an adaptive sequential replacement (ASR) method to improve the likelihood to achieve the best-fitting model with much reduced computational intensity. As detailed in Methodologies, we integrated stochasticity and adaptivity with the sequential replacement (SR) method to avoid local optimal solutions and unnecessary computational time when it is evident that a best-fitting model is achieved. The power for this ASR method was evaluated by simulated data. Our second objective was to compare the results between our ASR method and four other methods (SR, exhaustive, forward, and backward) with two actual genetic marker data sets. The purpose of this study is to provide a method to improve power to capture desirable genetic variation from high throughput genomic data for marker assisted selection with reduced computational intensity.

2. Methodologies

2.1. The ASR Algorithm

The SR procedure was detailed by Miller and usually converges rapidly. Unfortunately, this type of replacement algorithm does not guarantee convergence upon the best-fitting k-variable model [18]. In this study, we proposed the ASR algorithm to avoid local optimal solutions with a criterion to determine when the optimal solution is achieved, and the criterion used throughout this study is adjusted coefficient of determination, R_A^2 (or r-square for simplification). The

ASR procedure is detailed as follows:

Step 1: *Stochasticity process*. Randomly select a subset of k ($k \geq 2$) variables out of p candidate variables and set the variable index vector as id0. Run this k-variable linear regression analysis and calculate the r-square value as R_{A0}^2. This step focuses on stochasticity to avoid local optimal solution.

Step 2: *Sequential replacement process*. Replace the first variable in id0 with the remaining variables and run the k-variable linear regression analysis again and calculate the r-square value as R_{A1}^2 one by one with new variable index id1. If $R_{A1}^2 > R_{A0}^2$, set $R_{A0}^2 = R_{A1}^2$ and id0 = id1.

Step 3: Repeat step 2 for the second variable and the remaining variables in id0 if $k \geq 2$. Save R_{A0}^2 and id0.

Step 4: *Adaptivity process*. Repeat steps 1 to 3 until (1) the three largest r-square R_A^2 are identical, (2) the difference between the first and third largest adjusted R_A^2 is less than a given delta Δ (e.g. 0.001), or (3) it reaches a given maximum iteration time (e.g. 100) if condition (1) or (2) is not met. Save the largest r-square R_A^2 with the corresponding variables.

Step 5: Repeat steps 1 to 4 for N (*i.e.* 5 or 10) times. Record the largest r-square R_A^2 with the corresponding variable index vector.

Stochasticity is used in step 1 to avoid local optimal solutions. If condition (1) in step 4 is met, it is very likely that the optimal solution has been achieved. Step 5 will help increase the probability to reach the optimal if condition (1) is not met. If k is small less than 4 or the several candidate variables have a strong linear relationship with the response variable y, then the condition (1) in step 4 will be achieved rapidly. Given $p = 100$ and $k = 5$ the all-possible subset regression method, the number of linear regressions to be assessed is $C_p^k = \frac{p!}{k!(p-k)!} = 75287520$. While with our method, there are only $k*(p-k)+1 = 476$ multiple regression models to be assessed from steps 1 to 3. If step 4 is repeated for 50 times and step 5 is repeated for 5 times, the total number will be up to 119,000, which could be much less (0.16%) of computational time compared to the exhaustive method. In addition, either step 4 or 5 can be integrated with parallel computing to increase the computational speed proportionally to achieve the optimal solution.

2.2. Data Analysis

The authors of this study intended to compare the results between this ASR method and other commonly used methods. Such an intention is prohibited due to a few significant factors. For example, both forward selection and backward selection methods are available in several R packages but these two methods focus on selecting all significant rather than on k-variable model only. The exhaustive method is computationally prohibited for a large number of candidate variables. On the other hand, power and Type I error can be self-determined for a target method via simulation technique. Therefore, without losing focus, the authors of

this study emphasized on applying the ASR method to process the simulated data. While in applications, we aimed to compare the results among several methods: forward, backward, exhaustive, SR, and ASR methods. All four methods are available in the R package: leaps [14] [15] while the ASR method was developed by the first author of this paper and will be available upon request. In this study, all data simulations and actual data processing were conducted under RStudio platform [29] [30].

3. Results

3.1. Simulation Results

In our simulation study, a total of 100 independent variables ($p = 100$) were used while five ($k = 5$) were related to the response variable with equal contribution. The regression model used for simulation is as follows,

$$y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + b_4 X_{4i} + b_5 X_{5i} + e_i$$

where, y_i is response variable for observation i; b_0 is intercept, b_1 - b_5 are slopes for variables X_1 to X_5, respectively. For simplicity, the intercept and all slopes were preset to 1. Five sets of coefficients of correlation ($r = 0.00$, 0.20, 0.40, 0.60, and 0.80) among the first 10 variables are provided in **Table 1**, namely S1 ($r = 0.00$), S2 ($r = 0.20$), S3 ($r = 0.40$), S4 ($r = 0.60$), and S5 ($r = 0.80$). Four coefficients of determinations: $R^2 = 0.20$, 0.40, 0.60, and 0.80, equivalent to total heritability, from five variables/loci, were used. The above-mentioned parameters were used to generate simulated data. The mean power of five variables for each setting, mean adjusted coefficients of determination for selected $\overline{R}^2_{AS}$ and true models $\overline{R}^2_{AT}$ were calculated over 200 simulations.

Table 1. Coefficients of correlation between five true variables (X_1 - X_5) and other five noise variables (X_6 - X_{10}).

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.00	0.00	0.00	0.00	0.00	r	0.00	0.00	0.00	0.00
X_2	0.00	1.00	0.00	0.00	0.00	0.00	r	0.00	0.00	0.00
X_3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	r	0.00	0.00
X_4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	r	0.00
X_5	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	r
X_6	$r^\dagger$	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
X_7	0.00	r	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
X_8	0.00	0.00	r	0.00	0.00	0.00	0.00	1.00	0.00	0.00
X_9	0.00	0.00	0.00	r	0.00	0.00	0.00	0.00	1.00	0.00
X_{10}	0.00	0.00	0.00	0.00	r	0.00	0.00	0.00	0.00	1.00

†: $r = 0.00$, 0.20, 0.40, 0.60, and 0.80 and S1 - S5 were named accordingly.

The results are summarized in Table 2 and Table 3. The mean powers for the target variables being selected by our ASR method were 98.2%, 98.5%, 98.2%, 98.1%, and 96.5% for five settings S1, S2, S3, S4, and S5 when coefficient of determination was low as 0.20. When the coefficient of determination was 0.40 and higher, mean powers for target variables being selected was 100.0%. Therefore, the simulation results clearly suggest that this ASR method can be used to identify the best *k*-variable model, which can capture the maximum of variation in a linear regression analysis.

Comparing the coefficients of determination between the selected and the true models ($\overline{R}^2_{AS}$ vs $\overline{R}^2_{AT}$) helps us determine the efficiency of finding an optimal subset or better subset in linear regression analysis. The mean coefficients of determination for the models selected and the true models are summarized in Table 3. The results in Table 3 showed that mean R^2 for selected models was

Table 2. Mean powers of five quantitative variables being selected for five settings of correlation coefficients (0.00, 0.20, 0.40, 0.60, 0.80, S1, S2, S3, S4, and S5) among the first 10 variables and four coefficients of determination (R^2 = 0.20, 0.40, 0.60, and 0.80) each based on 200 simulated data sets.

Setting	Coefficient of determination			
	0.20	0.40	0.60	0.80
S1	0.982	1.000	1.000	1.000
S2	0.985	1.000	1.000	1.000
S3	0.982	1.000	1.000	1.000
S4	0.981	1.000	1.000	1.000
S5	0.965	1.000	1.000	1.000

Table 3. Mean adjusted coefficients of determination between selected models ($\overline{R}^2_{AS}$) and true models ($\overline{R}^2_{AT}$) over 200 simulations for five correlation settings (S1 - S5) with four different coefficients of determination (0.20, 0.40, 0.60, and 0.80).

		R^2			
		0.20	0.40	0.60	0.80
S1	$\overline{R}^2_{AS}$	0.1995	0.3995	0.6000	0.7979
	$\overline{R}^2_{AT}$	0.1992	0.3995	0.6000	0.7979
S2	$\overline{R}^2_{AS}$	0.2069	0.3978	0.5990	0.8014
	$\overline{R}^2_{AT}$	0.2065	0.3978	0.5990	0.8014
S3	$\overline{R}^2_{AS}$	0.2004	0.3976	0.5977	0.7974
	$\overline{R}^2_{AT}$	0.2000	0.3976	0.5977	0.7974
S4	$\overline{R}^2_{AS}$	0.2024	0.3969	0.5969	0.7991
	$\overline{R}^2_{AT}$	0.2022	0.3969	0.5969	0.7991
S5	$\overline{R}^2_{AS}$	0.2028	0.3955	0.5998	0.7989
	$\overline{R}^2_{AT}$	0.2020	0.3955	0.5998	0.7989

slightly higher than that for the original models when pre-set R^2 was 0.20. Checking individual R^2, we observed that each R^2 from each selected model was either equal to or higher than that for the true model (detailed results not provided). When R^2 was 0.40 or higher, R^2 for each selected model and that for the original model were identical for each simulated data set. The results in **Table 3** were highly consist with the those in **Table 2**. On one hand, when R^2 is small, occasionally, some true variables may be replaced by noise variables, which cause a slightly higher R^2 for that simulated data set due to Type I error. On the other hand, these results implied that this ASR method was able to identify a subset of variable with the highest R^2, which is desired mathematically.

3.2. Applications

In our first application, we applied the ASR method to a fruit fly wing data, which were used for QTL analysis [31]. The total number of polymorphic DNA markers on chromosome 2 is 37 ($p = 37$) after 11 co-existing markers were deleted. In this application, we were able to include SR, exhaustive, forward, and backward selection methods into our comparisons for $k = 1$ to 14. The SR, exhaustive, forward, and backward methods are available in leaps package [14]. The results in **Table 4** showed that R^2 for both exhaustive and ASR methods were identical, indicating that our ASR method has improved probability to determine the best k-variable/marker model. For most cases, the SR method had the same R^2 values compared to the ASR and exhaustive methods (*i.e.* $k = 1$ - 6, 8, 10, and 11) while the SR method had slightly lower R^2 values than the ASR and exhaustive methods for $k = 9$, 12, and 13 but not for $k = 7$ or 14. Both backward and forward selection methods had consistently and slightly lower R^2 value compared to the ASR and exhaustive selection methods except $k = 1$ for the forward selection method (**Table 4**). These two selection methods also yielded constantly lower R^2 values than the SR method for $k \geq 2$ except $k = 7$ and 14. The backward method had consistently higher R^2 values than the forward methods for all cases except for $k = 1$. It was surprising to notice that the R^2 values for $k = 7$ and $k = 14$ for SR method was far lower than those for the other four methods. We also noticed that the SR method yielded inconsistent and lower R^2 values for $k = 7$ and 14 when the order of 37 markers were randomized for several times (results not showed here). Without investigating the R scripts in the leaps package, it is hard to conclude if such outcomes were caused by the algorithm itself or bugs in leaps package.

In application 2, a barley data set, which was analyzed in our previous publication [32], was used to compare the SR, ASR, forward, and backward selection methods. The data set includes 391 single nucleotide polymorphisms (SNPs) and 762 heading data points. Due to high computational intensity, the analysis was prohibited by the exhaustive in leaps package; however, we were able to compare the results among four methods (SR, ASR, forward, and backward). The results showed that both SR and ASR methods performed equally well when $k = 1$ to 4,

6, and 7 while the ASR performed better for the remaining cases (Table 5). The forward method had higher R^2 values than the backward method except $k = 8$ and 9. Both SR and ASR methods had higher R^2 values than both forward and backward methods except $k = 1$ to 3 for the forward method, which had the same R^2 values compared to the SR and ASR methods.

Table 4. Adjusted coefficients of determination of k-marker subset ($k = 1$ to 14) for five selection methods for the data with fruit fly wing shape and 37 RFLP markers [31].

k	SR†	Exhaustive	ASR‡	Forward	Backward
1	0.514068	0.514068	0.514068	0.514068	0.445102
2	0.688750	0.688750	0.688750	0.642101	0.675081
3	0.843880	0.843880	0.843880	0.813288	0.836392
4	0.877729	0.877729	0.877729	0.867078	0.868122
5	0.903693	0.903693	0.903693	0.894649	0.897927
6	0.920823	0.920823	0.920823	0.911156	0.917549
7	0.230306	0.925672	0.925672	0.921786	0.922730
8	0.930037	0.930037	0.930037	0.926434	0.927707
9	0.931577	0.931974	0.931974	0.930089	0.930469
10	0.933271	0.933271	0.933271	0.931636	0.933069
11	0.934016	0.934016	0.934016	0.932646	0.933897
12	0.934320	0.934341	0.934341	0.933326	0.934223
13	0.934592	0.934617	0.934617	0.934060	0.934529
14	0.652253	0.934866	0.934866	0.934638	0.934806

†: sequential replacement and ‡: adaptive sequential replacement.

Table 5. Adjusted coefficients of determination of k-marker subset ($k = 1$ to 10) for four selection methods for the data with barley heading date and 391 SNPs [33].

k	SR†	ASR‡	Forward	Backward
1	0.318846	0.318846	0.318846	0.268571
2	0.371503	0.371503	0.371503	0.363640
3	0.407045	0.407045	0.407045	0.402651
4	0.434137	0.434137	0.432367	0.423071
5	0.450320	0.450409	0.450128	0.440955
6	0.467692	0.467692	0.462793	0.460830
7	0.483355	0.483355	0.475074	0.474516
8	0.492946	0.494132	0.486796	0.490362
9	0.501345	0.504865	0.498145	0.503484
10	0.513704	0.514244	0.508404	0.504725

†: sequential replacement and ‡: adaptive sequential replacement.

3.3. Repeatability, Consistence, and Speed

Repeatability and consistence for a method are important when stochasticity is applied to this method. The same data analyses in Applications 1 and 2 of this study with the ASR method were repeated independently for 20 times. The results including mean, minimum, and maximum of adjusted coefficients of determination for different *k*-marker models are summarized in Table 6 and Table 7, respectively. The results showed that the probability of the best model being determined for each *k*-marker set was at least 65% for the first application (Table 6) while the probability of the best model being determined varied widely among different *k*-marker sets for the second application (Table 7). The difference among mean, minimum, and maximum of adjusted coefficients of determination for each k-marker set was very small for both cases (Table 6 and Table 7). For example, the difference between minimum and maximum of adjusted

Table 6. Minimum, maximum, and mean values over 20 replications for R_A^2 for different k-marker models for the data with fruit fly wing shape and 37 RFLP markers [31].

	$k=1$	$k=2$	k = 3	$k=4$	$k=5$
Mean	0.514068	0.688750	0.843880	0.877727	0.903693
Min	0.514068	0.688750	0.843880	0.877675	0.903693
Max	0.514068(20†)	0.688750(20)	0.843880(20)	0.877729(19)	0.903693(20)
	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
Mean	0.920797	0.925672	0.930037	0.931974	0.933193
Min	0.920560	0.925672	0.930037	0.931974	0.932899
Max	0.920823(18)	0.925672(20)	0.930037(20)	0.931974(20)	0.933271(15)
	$k=11$	$k=12$	$k=13$	$k=14$	
Mean	0.934005	0.934332	0.934608	0.934864	
Min	0.933901	0.934286	0.934498	0.934856	
Max	0.934016(18)	0.934341(13)	0.934616(17)	0.934866(16)	

†: The number of the maximum R_A^2 was reached over 20 independent trials.

Table 7. Minimum, maximum, and mean values over 20 replications for R_A^2 for different k-marker models for the data with barley heading date and 391 SNPs [33].

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Mean	0.318846	0.371503	0.407045	0.433075	0.450381
Min	0.318846	0.371503	0.407045	0.432367	0.450128
Max	0.318846(20†)	0.371503(20)	0.407045(20)	0.434137(8)	0.450409(18)
	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
Mean	0.466777	0.481149	0.493443	0.504992	0.514079
Min	0.466675	0.477830	0.489611	0.502217	0.513864
Max	0.467692(2)	0.483355(10)	0.494132(14)	0.505904(5)	0.514667(5)

†:The number of the maximum R_A^2 was reached over 20 independent trials.

coefficients of determination was less than 0.0004 (equivalent to less than 0.040% lower than the exhaustive search) among 14 cases in application 1 (**Table 6**) and less than 0.006 (equivalent to less than 1.143% lower compared to the best model) in application 2. These results in **Table 6** and **Table 7** showed that the ASR method is repeatable and consistent in search of the best *k*-variable model. However, repeating the process from steps 1 to 4 for several times is recommended to reach better solutions for a large number of genetic markers which are closely linked as well.

The time used for selecting the best k-variable set will give us some insight in using the ASR method. The computer that we used for data processing was a Dell laptop with Intel® Xeon ® W-11855M CPU @ 3.20 GHz and 64.0 GB Ram. With the SR method, the time used in application 2 was less than 1 second for $k = 1$ to 10. With the ASR method, it averaged 28 minutes in total ($k = 1$ to 10) over 20 replications. Compared to the SR method, the ASR method is slow; however, this amount of time is very appealing because it can offer the improved power compared to the exhaustive method. With the ASR method, the given analysis for $k = 1$ to 10 in application 2 was completed within a lunchbreak period, which is very acceptable.

4. Discussion

Selecting a set of markers that captures the maximum genetic variation out of high throughput data is highly desired. The exhaustive search method is guaranteed to achieve the best solutions, but it can be prohibited due to high computational burden. On the other hand, many stepwise variable selection methods in linear regression analysis are heuristic and approximate, not guaranteeing the optimal solution, as showed in our two applications (**Table 4** and **Table 5**), though they offer desirable computational speed. The SR method starts a model selected from forward or stepwise selection method and then it allows to sequentially replace each variable in the model with the remaining variables [18] [34]. Mathematically, the SR method should be better than forward and backward selection methods. Therefore, the SR method has advantage of speed but it could result in a non-global optimal model due to lack of stochasticity. In this study, we were motivated to develop the ASR method with a high likelihood to identify a k-marker set for capturing the maximum genetic variation with much reduced computational intensity compared to the exhaustive method.

The first key feature applied to this ASR method is stochasticity to avoid local optimal solutions. Our ASR selection method starts a completely random *k*-variable subset every time as described in step 1. Thus, increasing the number of random starts should increase the possibility to avoid local optimal solutions and thus improve the possibility to reach global optimal solutions. However, such practice may add significant computational burden compared to the SR method. The other key feature added to the ASR method is adaptivity so that the search process can be terminated once the optimal solution is achieved as stated in con-

dition (1) at step 4. Our logic is that there could be several local optimal solutions for a *k*-variable model, but the global optimal solution is unique. If the best solution has appeared for at least three times during the search process, it is evident that the global solution has been achieved, and no additional search should be needed. For example, we preset 50 times of random start at step 4 but it can reach the global solution with only a minimum of five to 19 iterations for $k = 2$ to 14 in application 1, which greatly reduced computational intensity. However, users may need to increase the iteration number in condition (3) at step 4 or N at step 5 to increase the likelihood to achieve the best solution if condition (1) in step (4) is not met. In addition, this study showed that the ASR method is robust regarding obtaining highly repeatable and/or consistent best *k*-variable/marker models as showed in Table 6 and Table 7.

The power of this ASR method was numerically evaluated by simulated data through presetting five contributing variables at different levels of coefficients of determination, where each coefficient of determination is equivalent to heritability in genetics. Our simulation results showed that the ASR method was able to identify all optimal subset of true variables when the coefficient of determination was at least 0.40. Such a conclusion was evidenced by the coefficients of determination between the models selected and the true models and mean power over five preset variables being selected over 200 simulations (Table 3 and Table 4). Even when the preset coefficient of determination is low as 0.20 and target variables are highly correlated with noise variables (0.80), the ASR method sustained a high power as demonstrated in Table 2. In addition, we noticed that each individual R^2 values obtained by the ASR method was equal to or slightly greater than that for the true model for each simulated data set due to Type I error (Table 4, the individual results not provided here), indicating that it is more likely that the ASR method has the capability of selecting a better model. However, as expected, Type I error should be expected when a coefficient of determination or heritability is low.

Due to a small number of variables in our application 1, we compared five methods: forward, backward, SR, global, and our ASR methods. Both the ASR and exhaustive methods achieved the identical k-variable models ($k = 1$ - 14) (Table 4). The SR method could determine the same subset of variables for most cases when compared to the exhaustive method, indicating that the SR method has the ability to achieve the best model. However, occasionally, some models identified by the SR method had slightly lower coefficients of determination compared to those determined by the exhaustive and ASR methods. In two cases, the subsets determined by SR methods showed far lower coefficients of determination compared to the other four methods ($k = 7$ and 14 in the first application). The results showed that the SR method sometimes may lack consistence to generate the optimal solutions. In application 2, the ASR method had equal or higher adjusted coefficient of determination compared to the SR method for $k = 1$ to 10 (Table 5), suggesting that the ASR method sustains an im-

proved power and is preferred to identify the better models than the SR method.

Several key factors may influence the possibility to find the optimal model. The first factor is the degree of the subset linearly associated with the response/dependent variable. This is equivalent to heritability in a genetic association mapping study. Higher heritability is associated with higher power to catch the best model. The second factor is the degree of collinearity among variables associated with the dependent variable. Strong linear associations between predictive variables and the response variable and weak collinearity among predictive variables will achieve the optimal k-variable model much more rapidly than weak association and/or strong collinearity among the predictable variables. However, even though, increasing the number of random starts (iteration number) will help achieve the optimal solution and this is one desirable feature associated with this ASR algorithm, when the number of variables or genetic markers is high, more iterations are required. For example, over 100 iterations are more likely required to meet condition (1) or (2) in step 4 for $k \geq 6$ in the second application of this study.

Many association and QTL mapping studies showed that even though a single marker/locus was significantly associated with a quantitative trait of interest, using the single marker as MAS was still far from the efficiency needed for breeding selection. Thus, selecting k markers as a subset, which can catch desirable genetic variation, is desired for MAS application. However, it doesn't mean the more the better. In breeding practice, increasing one DNA marker for marker-assisted selection would double field/lab work with one additional bi-allelic DNA marker. On the other hand, our previous study on barley association mapping analysis showed that many selected SNP markers were significant yet the total coefficient of determination was stabilized with the increase of SNP markers at some points during our forwarded selection process [35]. The results from this study as presented in **Table 3** and **Table 4** also showed a similar pattern. Therefore, selecting a particular number of markers/variables should be determined depending on several key factors such as the degree of associations between selected genetic markers and the trait of interest and affordability/availability of labor and land. The ASR method in this study can help breeders capture the maximum genetic variation associated with a particular k-marker set.

The ASR selection method can potentially identify the best k-variable subset. However, it is possible that this method is extendable to forward and backward variable selections with slight modifications. For example, if all k variables in the model are significant, then steps 1 to 5 can be proceeded with $k + 1$ variables. This process can be repeated until no more new variables can be added. Such a process is related to the ASR based forward selection. On the other hand, if one or more variables are not significant in the k-variable solution, then steps 1 to 5 can be proceeded with $k - 1$ variables. This process continues until no more variables can be eliminated which is related to ASR backward selection. Additional comparisons between ASR based forward/backward selection methods and com-

monly used forward/backward selection are ongoing.

Acknowledgements

This study was partially supported by USDA-ARS (project # 6064-21000-016) and the USDA-NIFA hatch project (SD00H525-14) while the senior author formerly working at South Dakota State University.

Disclaimer

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Jansen, R.C. (1993) Interval Mapping of Multiple Quantitative Trait Loci. *Genetics*, **135**, 205-211. https://doi.org/10.1093/genetics/135.1.205

[2] Jansen, R.C. and Stam, P. (1994) High-Resolution of Quantitative Traits into Multiple Loci via Interval Mapping. *Genetics*, **136**, 1447-1455. https://doi.org/10.1093/genetics/136.4.1447

[3] Xu, S.H. (1995) A Comment on the Simple Regression Method for Interval Mapping. *Genetics*, **141**, 1657-1659. https://doi.org/10.1093/genetics/141.4.1657

[4] Zeng, Z.B. (1994) Precision Mapping of Quantitative Trait Loci. *Genetics*, **136**, 1457-1468. https://doi.org/10.1093/genetics/136.4.1457

[5] Zeng, Z.B. (2005) QTL Mapping and the Genetic Basis of Adaptation: Recent Developments. *Genetica*, **123**, 25-37. https://doi.org/10.1007/s10709-004-2705-0

[6] Haley, C.S. and Knott, S.A. (1992) A Simple Regression Method for Mapping Quantitative Trait Loci in Line Crosses Using Flanking Markers. *Heredity*, **69**, 315-324. https://doi.org/10.1038/hdy.1992.131

[7] Lander, E.S. and Botstein, D. (1989) Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics*, **121**, 185-199. https://doi.org/10.1093/genetics/121.1.185

[8] Hayes, B. (2013) Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). In: Gondro, C., van der Werf, J. and Hayes, B., Eds., *Genome-Wide Association Studies and Genomic Prediction*, Humana Press, Totowa, 149-169. https://doi.org/10.1007/978-1-62703-447-0_6

[9] Yu, J.M., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. and Buckler, E.S. (2006) A Unified Mixed-Model Method for Association Mapping That Accounts for Multiple Levels of Relatedness. *Nature Genetics*, **38**, 203-208. https://doi.org/10.1038/ng1702

[10] Berk, K.N. (1977) Tolerance and Condition in Regression Computations. *Journal of the American Statistical Association*, **72**, 863-866.

https://doi.org/10.1080/01621459.1977.10479972

[11] Gorman, J.W. and Toman, R.J. (1966) Selection of Variables for Fitting Equations to Data. *Technometrics*, **8**, 27-51. https://doi.org/10.1080/00401706.1966.10490322

[12] Efroymsn, M. (1966) Stepwise Regression—A Backward and Forward Look. Eastern Regional Meetings of the Institute of Mathematical Statistics, Florham Park, New Jersey.

[13] Draper, N. and Smith, H. (1966) Applied Regression Analysis. John Wiley & Sons, New York.

[14] Lumley, T. (2020) leaps: Regression Subset Selection. Version 3.1. https://CRAN.Rproject.org/package=leaps

[15] Hebbali, A. (2020) olsrr: Tools for Building OLS Regression Models. Version 0.5.3. https://CRAN.R-project.org/package=olsrr

[16] Venables, W.N. and Ripley, B.B. (2002) Modern Applied Statistics with S. Springer, New York. https://doi.org/10.1007/978-0-387-21706-2

[17] Hocking, R.R. (1976) The Analysis and Selection of Variables in Linear Regression (A Biometrics Invited Paper). *Biometrics*, **32**, 1-49. https://doi.org/10.2307/2529336

[18] Miller, A.J. (2002) Subset Selection in Regression. Chapman & Hall/CRC, Boca Raton.

[19] Mantel, N. (1970) Why Stepdown Procedures in Variable Selection. *Technometrics*, **12**, 621-625. https://doi.org/10.1080/00401706.1970.10488701

[20] Huberty, C.J. (1989) Problems with Stepwise Methods—Better Alternatives. *Advances in Social Science Methodology*, **1**, 43-70.

[21] Kirton, H.C. (1967) Best Models in Multiple Regression Analysis. N.S.W. Department of Agriculture, Sydney.

[22] Hocking, R.R. and Leslie, R.N. (1967) Selection of the Best Subset in Regression Analysis. *Technometrics*, **9**, 531-540. https://doi.org/10.1080/00401706.1967.10490502

[23] Beale, E.M.L., Kendall, M.G. and Mann, D.W. (1967) The Discarding of Variables in Multivariate Analysis. *Biometrika*, **54**, 357-366. https://doi.org/10.1093/biomet/54.3-4.357

[24] LaMotte, L.R. and Hocking, R.R. (1970) Computational Efficiency in the Selection of Regression Variables. *Technometrics*, **12**, 83-93. https://doi.org/10.1080/00401706.1970.10488636

[25] Furnival, G.M. and Wilson, R.W. (1974) Regression by Leaps and Bounds. *Technometrics*, **42**, 69-79. https://doi.org/10.1080/00401706.2000.10485982

[26] Miller, A.J. (1984) Selection of Subsets of Regression Variables. *Journal of the Royal Statistical Society Series A*, **147**, 389-425. https://doi.org/10.2307/2981576

[27] Barr, A.J., Goodnight, J.H. and Sall, J.P. (1979) SAS User's Guide. SAS Institute, Releigh.

[28] Lerner, J.V. and Games, P.A. (1981) Maximum R^2 Improvement and Stepwise Multiple Regression as Related to Overfitting. *Psychological Reports*, **48**, 979-983. https://doi.org/10.2466/pr0.1981.48.3.979

[29] R Core Team (2023) R: A Language and Environment for Statistial Computing. R Foundation for Statistical Computing, Vienna, Austria.

[30] R Studio Team (2022) R Studio: Integrated Development for R. R Studio, Inc., Boston.

[31] Weber, K., Eisman, R., Higgins, S., Morey, L., Patty, A., Tausek, M. and Zeng, Z.B.

(2001) An Analysis of Polygenes Affecting Wing Shape on Chromosome 2 in Drosophila Melanogaster. *Genetics*, **159**, 1045-1057.
https://doi.org/10.1093/genetics/159.3.1045

[32] Xu, Y., Wu, Y. and Wu, J. (2018) Capturing Pair-Wise Epistatic Effects Associated with Three Agronomic Traits in Barley. *Genetica*, **146**, 161-170.
https://doi.org/10.1007/s10709-018-0008-0

[33] Xu, Y., Bai, G.H., Graybosch, R., Wu, Y. and Wu, J. (2017) Marker Association Analysis with Three Agronomic Traits in Hard Winter Wheat Lines under Diverse Environments. *Journal of Applied Bioinformatics & Computational Biology*, **6**, 2.
https://doi.org/10.4172/2329-9533.1000136

[34] Myers, R.H. (1990) Classical and Modern Regression with Applications. PWS-KENT Publising Compary, Boston.

[35] Xu, Y., Wu, Y., Gonda, M. and Wu, J. (2015) A Linkage Based Imputation Method for Missing SNP Markers in Association Mapping. *Journal of Applied Bioinformatics & Computational Biology*, **4**, 1.