South Dakota State University
Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange

Animal Science Faculty Publications
Department of Animal Science

7-2015

A Linkage Based Imputation Method for Missing SNP Markers in Association Mapping

Yi Xu
South Dakota State University, yi.xu@sdstate.edu

Yajun Wu
South Dakota State University, yajun.wu@sdstate.edu

Michael G. Gonda
South Dakota State University, michael.gonda@sdstate.edu

Jixiang Wu
South Dakota State University, jixiang.wu@sdstate.edu

Xu et al., J Appl Bioinform Comput Biol 2015, 4:1
http://dx.doi.org/10.4172/2329-9533.1000115

Journal of Applied Bioinformatics & Computational Biology

Research Article

A SCITECHNOL JOURNAL

A Linkage based Imputation Method for Missing SNP Markers in Association Mapping

Yi Xu[1], Yajun Wu[2], Michael G Gonda[3] and Jixiang Wu[1,4*]

Abstract

This study proposed a novel linkage-based method for imputing missing DNA markers. This new method can be easily integrated with many other association mapping approaches to improve association mapping.

Association mapping has been widely used to detect desirable genetic markers associated with traits of interest for plant and animal improvement. Missing marker data area common and challenging issue in association mapping studies. Deleting individuals with missing markers can cause significant loss of important genetic information and lead to biased results and inappropriate conclusions. In this study, we proposed a linkage based imputation method for missing marker data given available linkage information. One significant advantage of this imputation method is its integrity with many currently available association mapping methods: once new data sets are imputed, many computer tools including various variable selection methods could be employed to determine markers associated with traits of interest. Imputation accuracy for this imputation method was evaluated by simulated data. As a demonstration, we applied this new approach to imputing missing data of single nucleotide polymorphism (SNP) markers in a barley data set and selected a set of SNP markers highly associated with heading date. Results showed that three of the five detected markers were associated with the regions or QTL of known of heading date control, suggesting that this new method is reasonably effective and robust in marker association study.

Keywords

Association mapping; Barley; Forward selection; Imputation; Heading date; SNP

Introduction

Association mapping has been widely used in detecting genetic markers associated with traits of importance in research areas such as plant breeding, human disease and animal breeding [1-6]. In recent years, various useful statistical methods and computing toolshave been developed for association mapping studies [7-12]. One of the critical challenges in association mapping is missing markers. For example, direct use of some of the statistical methods/tools mentioned above could be limited when some markers are missing. Therefore, it is helpful to fully use missing marker data in genetic mapping studies.

One commonly used method to deal with missing datais to remove the markers with any missing points, i.e. using only markers with complete data collection. List-wise (known as complete-case analysis) and pair-wise (known as available-case analysis) deletions are among the most common approaches of dealing with missing data [13]. Two major advantages of deletion methods are convenience and implementation speed. However, deletion methods may cause biased parameter estimation if the assumption of the "Missing Completely At Random" (MCAR) mechanismis not valid. The MCAR mechanism assumes that missing data are independent of other predictable variables, including the missing variable itself. Even if the MCAR assumption is valid, eliminating data can cause the loss of power and waste of information [14]. For single-marker analysis, it might be satisfactory to use deletion methods with a few missing data points; however deletion methods can be more problematic for multiple-marker analysis [15,14].

Instead of direct deletion of missing data, another commonly used method is imputation, replacing missing data with estimated values based on the observed data. Thus, the population size can remain the same with imputed data. Several imputation methods and tools based on a hidden Markov model (HMM) approach, such as IMPUTE [16,17] MACH [18,19], GERBIL [20], have been proposed. The key idea of HMM-based imputation methods is that haplotype are generated at random and then two haplotype are used to impute the missing genotypes [21]. In addition, some imputation methods like TUNA [22], SNPMSTAT [23] and PLINK [6] have been carried out based on SNP-tagging approaches [24]. Both HMM-based and SNP-tagging-based imputation methods are performed by using linkage dis equilibrium structure and reference datasets such as HapMap [25] in which a large set of SNPs are genotyped [21]. Unlike in human genotype imputation, it is sometimes difficult to find such a 'reference data set' in plant genotype imputations. Many of these methods are suitable for human rather than plant genotype imputations. For genotype imputation, commonly used software named TASSEL imputes missing markers using a k-nearest-neighbor algorithm [26].

Lander and Botstein developed interval mapping [27], which uses two flanking markers to determine each quantitative trait locus (QTL). In this study, the key idea of the interval mapping method was used to develop a new imputation method for missing DNA markers with linkage information available. This new method was evaluated by using simulated data. As a demonstration, we applied this method to a barley SNP data set with heading date. The reason to use barley heading date as our demonstration was that this trait is of agronomic importance and has been investigated. In addition, this imputation technique was also integrated with a forward selection method to identify DNA markers associated with heading date in barley.

Materials and Methods

Data collection

The phenotypic and genotypic data used in this study were initially downloaded from the Barley Coordinated Agricultural Project Hordeum Toolbox [28]. The cultivars used in this study included eight breeding groups developed by seven research institutions. Each group contained 96 lines (only 94 lines in the breeding group of University of Idaho). Table 1 showed that five groups were identified with over 2000 SNP markers; where as the other three groups were identified with over 1200 SNP markers but fewer than 2000. AA, AB and BB,

*Corresponding author: Jixiang Wu, Department of Plant Science, Box 2140C, South Dakota State University, Brookings, SD 57007, USA, Tel: 605-688-5947; E-mail: jixiang.wu@sdstate.edu

Received: March 16, 2015 Accepted: July 17, 2015 Published: July 24, 2015

Table 1: Summarized information of SNP marker data sets in eight data groups[a]

\Data Group in Institution	Number of Lines	Number of SNPs	Proportion (%) of Genotype			
			AA	BB	AB	MM[b]
University of Idaho	94	2324	48.53	50.42	0.62	0.42
Busch Agricultural Resources Inc	96	2312	47.98	51.58	0.12	0.33
University of Minnesota	96	1290	49.87	49.60	0.33	0.20
Montana State University	96	1537	50.27	48.98	0.25	0.49
North Dakota State University(NDSU 2-Row)	96	2273	48.33	50.36	0.67	0.64
North Dakota State University(NDSU 6-Row)	96	1333	49.71	49.43	0.47	0.38
Utah State University	96	2489	38.25	46.92	0.98	13.85
Washington State University	96	2335	51.49	46.86	0.88	0.77

Note: [a]Information collected from http://hordeumtoolbox.org
[b]Indicates the proportion of missing markers

where AA and BB are homozygous and AB heterozygous expressed the SNP markers as three genotypes, which were denoted. Except for the breeding group of Utah State University, proportions for genotypes AA and BB were ~50% each while the proportion for the heterozygous genotype AB was less than 1% (Table 1). Although the breeding group of Utah State University contained13.85% missing markers, which were relatively higher than others groups, its heterozygous rate was lower than 1%. We identified 448 SNP markers that were commonly shared in all genotypes. Among these 416 SNPs were distributed on seven chromosomes with linkage information available, ranging from 0 to 0.33. However, thirty two SNP markers were removed because of missing linkage information. Since all these lines were selfed at least to the F_5-generation [29], the heterozygous genotype rate (AB) was slow (0.72%) among these 416 SNP markers, it is reasonable to assume that these missing markers are more homozygous. As a demonstration of the use of this method, only one important agronomic trait, heading date measured intwo environments (with and without irrigation conditions), was used in this study. On the other hand, the genotype-by-environment interaction effects had little impact on this trait as reported previously [29], thus, the mean values over two environments were used for association mapping in this study. The missing rates for these SNPs are summarized in Table 2.

Genotype imputations

Unlike imputation for quantitative variables, the key idea of genotype imputation used in our study is that a missing SNP marker can be considered as a binary random variable [30], whose probability distribution can be derived based on the allelic information of one or two flanking markers as provided in this study. Then, each missing marker can be sampled based on a derived probability distribution.

There are two general cases of missing genotypes. The first case is that there are two flanking markers for a missing locus and the second case is that there is only one flanking marker for a missing locus. Derivations of the probability for these two cases are detailed as following.

For the first case with two flanking loci, the three markers were denoted as locus 1 (with two alleles A and a), 2, and 3 (with two alleles B and b), where locus 2 has unknown alleles M or m) with r_1 recombination fraction from locus 1 and r_2 recombination fraction from locus 3. The flanking loci were locus 1 and locus 3 with recombination fraction *r*. The relationship among recombination fractions is $r = r_1 + r_2 - 2r_1r_2$ [31]. When r is small, no double crossover can be assumed and the equation is equivalent to $r = r_1 + r_2$ [31]. Considering the presence of double crossovers, the expected probability of AABB could be written as a function of r: P (AABB)=0.5(1-r) and the expected probabilities of AAMMBB and AAmmBB could be written as follows: P (AAMMBB) =0.5(1-r_1) (1-r_2) and P (AAmmBB)=0.5 r_1r_2, respectively. Using the same principle, the expected probabilities for other genotypes could be derived. If A and B are tightly linked, then the possibility of double crossover could be ignored, so that the cases of AmB and aMb would be very rare [31]. As a result, the expected probability of AAmmBB or aaMMbb could be close to zero. The detailed results of expected genotype probabilities assuming the presence or absence of double crossovers are listed in Tables 3 and 4 respectively. In Rubin's imputation rule, the proportion of missing genotypes is a variable with a probability distribution [14,30], so in the next step the probability distribution of the missing SNP marker based on flanking SNP markers needs to be derived. To achieve this goal, we derived an expected probability of a missing SNP marker conditional on flanking SNP markers by the conditional probability equation such as:

$$P(MM \mid AABB) = \frac{P(AAMMBB)}{P(AABB)} = \frac{(1-r_1)}{1-r}\frac{(1-r_2)}{1-r}$$

and or the second case where there is only one flanking marker available, the probability of a missing SNP marker conditional on two flanking SNP markers introduced previously is not needed because only one flanking marker, either locus 1 or 3, is available. The solution for this case is to use the information of one flanking SNP marker to calculate the conditional probabilities. For example, the expected probability of genotype "AAMM" (no flanking marker genotypeon locus 3) is 0.5 (1-r_1). Similarly, the expected probabilities of a missing SNP marker conditional on one flanking SNP marker are $P(MM \mid AA) = \frac{P(AAMM)}{P(AA)} = 1 - r_1$ and $P(mm \mid AA) = \frac{P(AAmm)}{P(AA)} = r_1$.

In some cases, recombination fractions need be estimated and converted from genetic distances. Several recombination fraction estimation methods exist [32-36]. In this study, though various mapping functions can be used, we used the Haldane's map function [33] to estimate the recombination fraction: $r = 0.5(1 - e^{-2|d|})$, where d is defined as the map distance between two marker loci. The users may use other mapping functions too.

Once the conditional probability of a missing marker is derived, we can impute each missing marker. After all missing markers are imputed; a new data set could be generated. Then this new data set could be used for further association mapping analyses with various statistical methods and software [12,10,7, 8, 11,9]. Via this algorithm, new data sets could be repeatedly generated and analyzed.

Design of simulations

To validate the proposed genotype imputation method, simulation studies were designed for both cases (with two flanking markers and one flanking marker). For each simulated data set, we generated 10,000 individuals each with three bi-allelic marker loci. The three markers were denoted as locus 1 (with two alleles A and a), 2 (with two alleles M and m) and 3 (with two alleles B and b). Locus 2 was in the middle with r_1 recombination fraction from locus 1 and r_2 recombination fraction from locus 3. Without loss of generality, we assumed $r_1 \leq r_2$ in all our simulation studies to reduce computation demands. Marker information for locus 2 was considered missing. For case 1 we only used marker information from loci 1 and 3, while for case 2 we only used the information form locus 1. The imputation accuracy over all missing points for each data set was calculated (correct number of imputed genotypes/total missing points). Mean accuracy estimates with corresponding standard deviations (SD) were obtained from 100 simulated data sets based on different fixed preset recombination fractions (r_1 and r_2) and are reported in Tables 5 and 6.

Association mapping approaches

In an actual data analysis, in addition to dealing with missing genotypic data, researchers are interested in detecting a group of markers associated with a target trait. Once an imputed data set is generated, a set of markers associated with a quantitative trait could be determined by using the following multiple linear regression models:

$$y_i = \mu + \sum_{j=1}^{p} b_j x_{ij} + e_i \; (i = 1, 2, ..., n; \; j = 1, 2, ..., p) \qquad (1)$$

Where y_i is the phenotypic value of the ith line; n is the number of lines; μ is the intercept; b_j is the effect of marker j; p is the number of causal loci; x_{ij} is an indicator variable that takes a value of 0 or 1 if the genotype of the ith line at marker j is AA or BB, respectively; and $e_i \sim N(0, \sigma_e^2)$ is a random error. In this study, we assumed all marker effects were fixed with main effects only.

As mentioned above, a number of successfully used statistical methods and software such as LASSO [7-9] and random forest [8,10] could be employed for determining a set of markers associated with a trait. In order to compare the results of association mapping between pair-wise deletion and linkage based imputation methods, in this study, we focused on a forward selection. By doing so, two criteria were usedwith forward selection: adjusted coefficient of determination (R^2) and the *p-value.*

The forward selection method was used with both deletion methods and the imputation proposed in this study. With deletion methods, all individuals with one or more missing markers would be deleted before forward selection is applied. In this study, in order to maximize the use of marker information, we used a pair-wise deletion method, which deleted the individuals with missing value (s) when only markers were included in the model. By doing so, the data sizes remain the same for an imputed data set while it could decrease as the number of markers included in the model increases with the use of the deletion approach. With the use of this linkage based imputation method, imputed data could be repeatedly generated and analyzed. In this study, imputed data sets were repeatedly generated 1,000 times, then the frequency of each marker being selected was calculated. The number of markers selected in the model among these imputed data set varied from 6 to 16, where we observed that the increase in R^2

Table 2: Summarized information of SNP marker missing rate for each chromosome.

Chromosome	Number of SNPs	Marker Missing Rate (%)		
		Minimum	Maximum	Mean
1	36	0.13	8.27	2.50
2	59	0.26	9.45	2.75
3	78	0	33.33	2.64
4	47	0	13.65	3.39
5	102	0	18.64	3.14
6	52	0.26	8.79	3.34
7	42	0.13	33.73	3.69
Total	416	-	-	3.04

Table 3: Expected genotypic probability for missing marker (MM/mm) and two flanking markers (AA/aa and BB/bb) with double crossover.

Flanking Marker Genotype	Probability for Flanking Markers	Probability	
		MM	mm
AABB	$0.5(1-r)$	$0.5(1-r_1)(1-r_2)$	$0.5r_1r_2$
AAbb	$0.5r$	$0.5(1-r_1)r_2$	$0.5r_1(1-r_2)$
aaBB	$0.5r$	$0.5r_1(1-r_2)$	$0.5(1-r_1)r_2$
aabb	$0.5(1-r)$	$0.5r_1r_2$	$0.5(1-r_1)(1-r_2)$

Table 4: Expected genotypic probability for missing marker (MM/mm) and two flanking markers (AA/aa and BB/bb) with no double crossover.

Flanking Marker Genotype	Probability for Flanking Markers	Probability	
		MM	mm
AABB	$0.5(1-r)$	$0.5(1-r)$	0
AAbb	$0.5r$	$0.5r_2$	$0.5r_1$
aaBB	$0.5r$	$0.5r_1$	$0.5r_2$
aabb	$0.5(1-r)$	0	$0.5(1-r)$

Table 5: Imputation accuracy estimates averaged for 100 simulations based on fixed recombination fractions with double crossovers (two flanking markers).

r_2 r_1	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0	1(0)[a]	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)
0.05		0.9951(0.0007)	0.9901(0.0009)	0.9852(0.0013)	0.9802(0.0013)	0.9754(0.0015)	0.9705(0.0017)	0.9661(0.0019)	0.9616(0.0016)	0.9565(0.0021)	0.9525(0.0021)
0.10			0.9807(0.0013)	0.9704(0.0018)	0.9610(0.0021)	0.9518(0.0022)	0.9426(0.0024)	0.9341(0.0024)	0.9256(0.0024)	0.9175(0.0027)	0.9102(0.0027)
0.15				0.9565(0.0020)	0.9429(0.0023)	0.9290(0.0025)	0.9163(0.0031)	0.9040(0.0028)	0.8924(0.0033)	0.8819(0.0033)	0.8721(0.0030)
0.20					0.9249(0.0026)	0.9077(0.0028)	0.8917(0.0030)	0.8767(0.0034)	0.8622(0.0033)	0.8510(0.0036)	0.8401(0.0038)
0.25						0.8871(0.0029)	0.8696(0.0030)	0.8518(0.0032)	0.8356(0.0035)	0.8240(0.0039)	0.8130(0.0036)
0.30							0.8477(0.0037)	0.8289(0.0040)	0.8139(0.0034)	0.7997(0.0040)	0.7901(0.0039)
0.35								0.8099(0.0044)	0.7941(0.0042)	0.7809(0.0044)	0.7724(0.0038)
0.40									0.7786(0.0040)	0.7666(0.0041)	0.7592(0.0039)
0.45										0.7572(0.0047)	0.7525 (0.0041)
0.50											0.7498(0.0039)

Note: [a]The value in round brackets are standard deviations from 100 simulated data sets.

Table 6: Accuracy estimates averaged for 100 simulations based on fixed recombination fraction (one flanking marker).

r	**0**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
Accuracy Estimate	1(0) [a]	0.9052(0.0027)	0.8195(0.0040)	0.7452(0.0049)	0.6805(0.0047)	0.6252(0.0046)	0.5798(0.0040)	0.5445(0.0052)	0.5200(0.0047)	0.5047(0.0048)	0.5002(0.0049)

Note: aThe value in round brackets is standard deviation from100 simulated data sets.

started stabilizing after five markers were selected for each data set (Table 7 and Figure 1). Thus, we only focused on the frequencies of the first five markers being selected into the models.

The forward selection method was used to select a group of markers with maximum contribution to heading date in barley. For comparison both imputation and deletion methods were considered and integrated with the forward selection method. When using the imputation method, for each imputation data set, we denoted the cumulative adjusted coefficient of determination as CR_I^2, (k=1,2,…,1000) and the mean imputation-based cumulative adjusted coefficient of determination as CR_I^2 following the equation:

$$CR_I^2 = \frac{1}{1000}\sum_{K=1}^{1000} CR_{I(K)}^2.$$

When using the deletion method, we applied forward selection (selection step was fixed at five) based on pair wise deletion methods and the corresponding cumulative adjusted coefficient of determination was denoted as CR_D^2. Since the cumulative adjusted coefficient of determination were stabilized after five markers being selected in each model, we only compared the results of five-marker models for two methods.

All data analyses and computations were conducted by R codes developed by the authors in this study under the R platform (Version 3.0.1) and the authors of this paper wrote the code. An Rpackage'linkim' for the proposed imputation method is available online [37].

Results

Simulation results

The imputation accuracies for different combinations of recombination fractions are summarized in Table 5 (with two flanking markers) and Table 6 (with one flanking marker), respectively. The results in Table 5 showed that estimated accuracy increased as r_1 decreased. For example, the accuracy rate was 0.9052 for r_1 equal to 0.05 with one flanking marker. With two flanking markers available, the estimated accuracy ranged from ~75% to 100% (Table 5). For example, when r_1=0005, the estimated accuracy was greater than 95%. Generally, if one of r_1 and r_2 was fixed, the accuracy increased as the other recombination fraction decreased. With one flanking marker available, the estimated imputation accuracy ranged from ~50% to 100% (Table 6). The above results indicated that using two flanking markers could improve imputation rate as compared to using one flanking marker.

As for the barley SNP marker data used in this study, there were a total of 9,643 missing genotype points (equivalent to ~ 3.4% of missing markers). Among these missing genotype points, 96.34% had two flanking markers (recombination fractions r_1 and r_2 ranged from 0 to 0.19 and from 0 to 0.33, respectively) while another 3.66% had one flanking marker (recombination fraction r_1 ranged from 0 to 0.17). Based on the simulation results (Tables 5 and 6), the estimated mean accuracy of the imputation data for this barley SNP marker data set was greater than 0.95.

SNPs associated with heading date

Once all missing SNP marker data were imputed, we used a forward selection method to determine SNPs associated with barley heading date. Since we observed that the adjusted coefficient of determination R^2 was stabilized after five SNP markers were included in each model for each imputed data set, we focused on the first five (i.e.p=5) selected SNP markers associated with heading date for each imputed data set (Table 7). Of these five SNPs, SNPs *11_10262* and *11_20868* were located in genomic regions known to be associated with heading date and SNP *11_11002* was located within 6 cMofa known quantitative trait locus (QTL) for heading date on chromosome 3 [38]. Our analyses showed that the probability ofSNP *11_11002*(chromosome 3) being selected into the model was 99.9% at the first step, followed by SNP *11_10262* (chromosome 4) with a 100%probability of being selected. The other three markers (SNPs*11_10551*,*12_31469*and *11_20868*) were selected with almost the identical probability of 87.1% (Table 7 and Figure 1). The adjusted coefficient of determination (R^2) for these five SNP markers in the model was 44.98% while the first SNP (SNP *11_11002*) contributed to 31.59% of the total variation in heading date. No SNP markers selected were located on chromosomes 1 or 5 (based on 1,000 imputed data sets in this study) where as only one SNP marker was identified on chromosome 7 with a probability<15% (Figure 1) [39-44].

Comparison of imputation and deletion methods

For comparison of imputation and deletion methods, we considered the cumulative adjusted coefficient of determination based on the first five markers as a criterion. We observed that the imputation-based cumulative R^2 (CR_I^2) (Table 7) was slightly smaller than the deletion-based cumulative R^2(CR_D^2) (Table 8). One possible reason was that the reduced sample sizes via deletion methods could cause a slightly higher R^2. In addition, SNPs *11_11002*, *11_10262* and *12_31469* were detected based on both imputation and deletion methods, but SNPs *11_20868* and *11_10551* were only detected based on our imputation method while SNPs *11_21209* and *12_30691* were only detected based on the deletion method.

Discussion

In order to improve mapping power and the utilization of genetic marker data, compute tools are needed to deal with missing marker data. Several genotype imputation methods were developed to impute human genotypes [16-22,44-46] based on a reference data set. In plant genotype imputations, such a 'reference data set' may not be available, yet linkage information could be easily obtained when linkage maps

Table 7: Summary of detected SNP markers associated with heading date of barley detected in 1,000 imputations usinga forward selection method.

Step	SNP	Chr	Position (cM)	Selected marker frequency (%)	CR_I^2	Missing Rate (%)	Cumulative Missing Rate (%)
1	11_11002	3H	43.99	99.9	0.3159	3.81	3.81
2	11_10262	4H	55.63	100.0	0.3696	2.23	5.25
3	11_10551	2H	139.65	87.1	0.4062	2.89	7.09
4	12_31469	6H	126.18	87.1	0.4319	6.96	12.99
5	11_20868	6H	124.85	87.1	0.4498	4.46	14.83

Abbreviations: SNP=Single Nucleotide Polymorphism; Chr = chromosome number; CR_I^2 = Cumulative adjusted coefficient of determination based on imputation method.

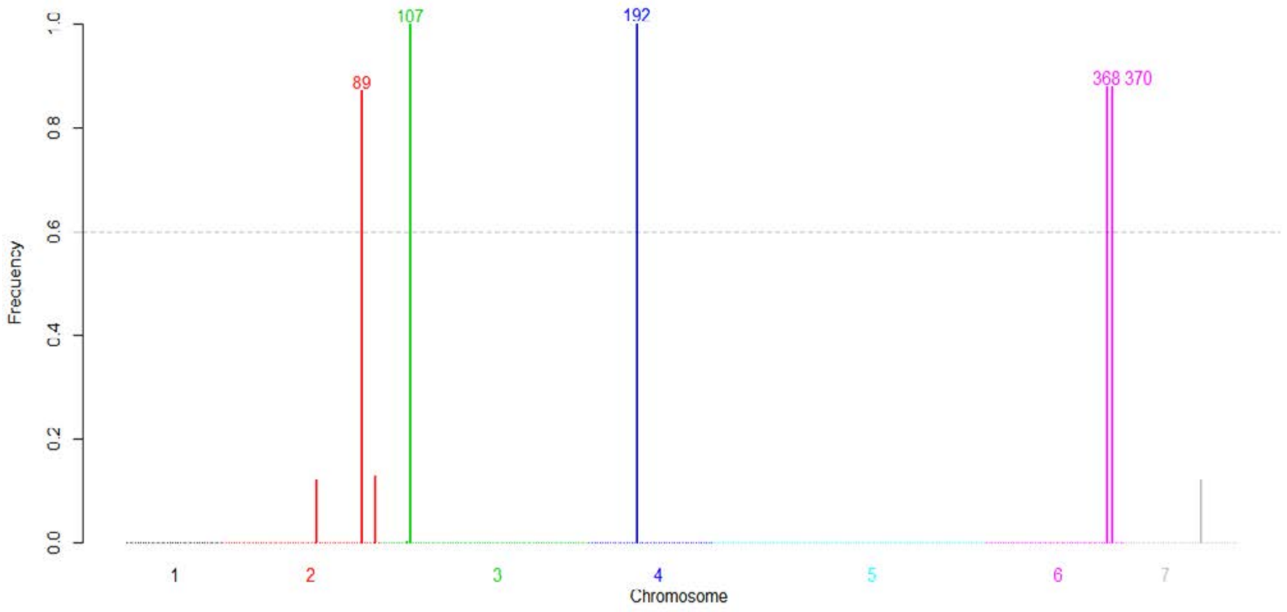

Figure 1: The estimated frequencies for the selected SNP markers using forward selection for1000 imputations
Abbreviations: 89=SNP11_10551; 107= SNP 11_11002; 192= SNP 11_10262; 197= SNP 12_30237; 368=SNP 11_20868; 370=SNP 12_31469.

Table 8: Summary of detected SNP markers associated with heading date of barley detected using forward selection based on the deletion method.

Step	SNP	Chr	CR_D^2	Missing Rate (%)	Cumulative Missing Rate (%)
1	11_11002	3H	0.3207	3.81	3.81
2	12_31469	6H	0.3703	6.96	10.10
3	11_10262	4H	0.4057	2.23	11.42
4	11_21209	7H	0.4399	3.28	13.12
5	12_30691	2H	0.4597	9.45	16.53

Abbreviations: SNP=Single Nucleotide Polymorphism; Chr = chromosome number; CR_D^2 = Cumulative adjusted coefficient of determination based on deletion method.

for different populations are established. In this study, we proposed a new approach to impute missing marker data based on available linkage information. It could be an important add it into the current methods for association mapping studies.

The simulated results showed that our imputation method provided high statistical accuracy in the case of two flanking markers (Table 5). The estimated accuracy was higher if the missing locus was more tightly linked with flanking markers. For the case of one flanking marker, the estimated accuracy was from 0.5 to 0.9, if the recombination fraction was between 0.5 and 0.05 (Table 6). Normally for the same recombination fraction, the estimated accuracy for the case where two flanking markers are available is greater than that for the case where one flanking marker is available. The standard deviations of accuracies in Tables 5 and 6 were smaller than 0.01, indicating that the proposed imputation method was stable based on the simulations.

Regarding the actual barley data, we not only imputed the missing SNP data, but also integrated a forward selection method to identify a group of markers associated with heading date (Table 7). In this study, five important SNP markers associated with barley heading date were detected. These five SNP markers contributed about 45% of the total variation in heading date.

Once a new data set is generated by imputation methods, many computer tools could be employed for association mapping studies. In this study, we have shown how thisimputation method could be integrated with a forward selection method to identify a set of markers associated with a quantitative trait of interest. It should be pointed out that many other variable selection methods could be employed as well. Our results showed that the first five markers were consistently selected (with high percentage being selected) based on 1,000 imputed data sets (Table 7 and Figure 1). We also observed that the selected SNP *11_11002* located on chromosome 3 contributed the highest heritability for heading date, and previous studies showed that there was a QTL near this SNP within 6 cm [38] associated with heading date in barely.

The results obtained from genetic association mapping could be highly impacted by high missing data rate [39], especially with the use of the deletion methods. One advantage of the imputation method has over the deletion methods is that the sample size remains the same by imputing the missing data so that statistical power for association mapping could be improved compared to the deletion methods [40]. For example, the imputation method will be useful to identify DNA markers associated with several traits simultaneously, which might be impossible after the deletion of markers due to missing marked data. The second advantage of using imputation methods is the possible integration with many association mappingapproaches. The R package, 'linkim', we developed in this study could be easily integrated with many other variable selection methods like LASSO [7-9] and random forest [10]. In addition, we found that the results of this barley data set by using the proposed imputation method were similar to the results by a frequency based-single marker imputation method [41]. One reason to explain the similar results from these two imputation methods might be the low missing rate in this barley data set.

In this study, we observed that the cumulative adjusted coefficient of determination for both imputation data and deletion data were stabilized after five markers were selected by the forward selection method and the identified markers were similar. Three markers(SNPs *11_11002*, *11_10262* and *12_31469*) were detected by both imputation and deletion methods; however, the orders that these markers being selected in the models were different for these two methods (Tables 7 and 8). The total contribution of the selected markers to barley heading date for the two methods was similar; however, the deletion method only used 126 genotypes (16.53% out of 762 genotypes in the data set) were deleted when these five markers were selected (Table 8). To demonstrate the effect of missing marker data on the association analysis, the barely data was deleted randomly to create10% and 15% of missing rates and the modified data sets were then subjected to the deletion and the imputation methods for association analysis. The 10% and 15% of missing rates led to more than 30% and 50%, respectively, genotype removal from the analysis if using the deletion method, resulting in ~3% and 22%, respectively, of chance that at least one of the five marker selected by the imputation method cannot be significantly detected- 24% and 66%, respectively, of chance that at least one of the five markers selected from the original data set using the deletion method cannot be detected. On the other hand, the use of the imputation method on the data with 10% and 15% of missing rate always (100%) led to the identification of the same five markers selected from the original data set using the imputation method. These results strongly suggest that that our imputation based method yields more consistent and thus potentially more reliable marker selection compared to the deletion method. This conclusion is consisted with the findings [14]. Therefore, the proposed imputation method is preferred over the deletion method, especially for the data set with a high rate of missing markers.

Though barley SNP marker data were used in this study, the proposed approach could be applied to other types of genetic markers once linkage information is available. Also, the proposed imputation method could be extended to heterozygous genotypes. By modifying our genotype imputation methods, we developed an R package 'linkim' [42] that is available on the Comprehensive R Archive Network (CRAN) website[43].

Acknowledgments

This work waspartially supportedby USDA-NIFA Hatch project --1005459, South Dakota State University Research Scholarship Support Fund Project, and the South Dakota Agricultural Experiment Station at South Dakota State University.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Donnelly P (2008) Progress and challenges in genome-wide association studies in humans. Nature 456: 728-731.
2. Hayes BJ, Bowman PJ, Chamberlain AJ, Savin K, Van Tassell CP, et al. (2009) A validated genome wide association study to breed cattle adapted to an environment altered by climate change. PLoS One 4: e6676.
3. Hayes B, Goddard M (2010) Genome-wide association and genomic selection in animal breeding. Genome 53: 876-883.
4. Herzog T, Rubin DB (1983) Using multiple imputation to handle non response in sample surveys. Academic Press, Inc, New York. USA.
5. Nicolae DL (2006) Testing untyped alleles (TUNA)-applications to genome-wide association studies. Genet Epidemiol 30: 718-727.
6. Wang D, Zhu J, Li Z, Paterson A (1999) Mapping QTLs with epistatic effects and QTL×environment interactions by mixed linear model approaches. Theor Appl Genet 99: 1255-1264.
7. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32: 407-451.
8. Cherkassky V, Ma Y (2009) Another look at statistical learning theory and regularization. Neural Netw 22: 958-969.
9. International HapMap Consortium (2003) The international hap map project. Nature 426: 789-796.
10. Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, et al. (2006) SNP-based analysis of genetic substructure in the German population. Hum Hered 62: 20-29.
11. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106: 9362-9367.
12. Breiman L (1984) Classification and regression trees. Wadsworth International Group, Belmont, California, USA.
13. Lin DY, Hu Y, Huang BE (2008) Simple and efficient analysis of disease association with missing genotype data. Am J Hum Genet 82: 444-452.
14. Enders CK (2010) Applied missing data analysis. Guilford Press, New York, USA.
15. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7: 781-791.
16. Holland JH (1992) Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence. MIT Press, Cambridge, Mass, London, UK.
17. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11: 499-511.
18. Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185-199.
19. Shahinnia F, Rezai A, Sayed-Tabatabaei BE, Komatsuda T, Mohammadi SA (2006) QTL mapping of heading date and plant height in Barley cross "Azumamugi" × "Kanto Nakate Gold". Iran J Biotech 4: 88-94.
20. Karlin S (1984) Theoretical aspects of genetic map functions in recombination processes in human population genetics: The Pittsburgh Symposium, New York, USA.
21. Halperin E, Stephan DA (2009) SNP imputation in association studies. Nat Biotechnol 27: 349-351.
22. Newman DA (2003) Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. Organizational Research Methods 6: 328-362.
23. Liang M, Hole D, Wu J, Blake T, Wu Y (2012) Expression and functional analysis of nuclear factor-y, subunit b genes in barley. Planta 235: 779-791.
24. Liu BH (1998) Statistical genomics: linkage, mapping, and QTL analysis. CRC Press LLC, USA.
25. Sun YV, Kardia SL (2008) Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. Eur J Hum Genet 16: 487-495.
26. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, et al. (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635.
27. Lachenbruch PA (2011) Variable selection when missing values are present: a case study. Stat Methods Med Res 20: 429-444.
28. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42: 355-360.

29. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10: 387-406.

30. Dean CB, Nielsen JD (2007) Generalized linear mixed models: a review and some extensions. Lifetime Data Anal 13: 497-512.

31. Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York, USA.

32. Felsenstein J1 (1979) A mathematically tractable family of genetic mapping functions with different amounts of interference. Genetics 91: 769-775.

33. Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8: 299-309.

34. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. Plos Genet 5: e1000529.

35. Korte A, Vilhjalmsson BJ, Segura V, Platt A, Long Q, et al. (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44: 1066-1071.

36. Miller AJ (2002) Subset selection in regression. Chapman & Hall/CRC, Boca Raton, FL, USA.

37. http://hordeumtoolbox.org.

38. Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3: e114.

39. Stich B, Möhring J, Piepho HP, Heckenberger M, Buckler ES, et al. (2008) Comparison of mixed-model approaches for association mapping. Genetics 178: 1745-1754.

40. Browning SR (2008) Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet 124: 439-450.

41. Kosambi DD (1943) The estimation of map distances from recombination values. Ann Eugen 12: 172-175.

42. http://cran.r-project.org/web/packages/linkim/index.html.

43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

44. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81: 1084-1097.

45. Pei YF, Li J, Zhang L, Papasian CJ, Deng HW (2008) Analyses and comparison of accuracy of different genotype imputation methods. PLoS One 3: e3551.

46. Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, et al. (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44: 825-830.

Author Affiliation

Top

[1]Department of Plant Science,South Dakota State University, Brookings, SD 57007, USA
[2]Department of Biology and Microbiology,South Dakota State University, Brookings, SD 57007, USA
[3]Department of Animal Sciences,South Dakota State University, Brookings, SD 57007, USA
[4]Department of Mathematics and Statistics,South Dakota State University, Brookings, SD 57007, USA