Published March, 2009

Marker Imputation in Barley Association Studies

Jean-Luc Jannink,* Hiroyoshi Iwata, Prasanna R. Bhat, Shiaoman Chao, Peter Wenzl, and Gary J. Muehlbauer

Abstract

Association mapping requires high marker density, potentially leading to many missing marker data and to high genotyping costs. In human genetics, methods exist to impute missing marker data and whole markers typed in a reference panel but not in the experimental dataset. We sought to determine if an imputation method developed for human data would function effectively in a barley (*Hordeum vulgare* L.) panel. The panel contained 98 lines, 2517 single nucleotide polymorphism (SNP) markers, and 716 Diversity Arrays Technology (DArT) markers. Averaged over markers, masked scores were correctly imputed 97.1% of the time. We chose 610 and 273 tag markers in two- and six-row barley subpopulations, respectively. Despite this low number of tags, imputation accuracy was such that for about 80% of non-tag markers, the prediction r^2 between imputed and true scores was 0.8 or higher. When DArT markers were used as tags, SNP markers were imputed with similar accuracy, suggesting that the method can convert association information from one marker system (e.g., DArT) to another marker system (e.g., SNP). We believe marker imputation methods will have an important future in association studies as a component of tagging methods and in reducing problems due to missing data.

Published in The Plant Genome 2:11–22. Published 13 Feb. 2009.
doi: 10.3835/plantgenome2008.09.0006

677 S. Segoe Rd., Madison, WI 53711 USA
An open-access publication

THE OBJECTIVE OF GENETIC mapping is to identify simply inherited markers in close proximity to genetic factors affecting quantitative traits (quantitative trait loci, or QTL). This localization relies on processes that create a statistical association (called linkage or gametic phase disequilibrium and henceforth abbreviated LD) between marker and QTL alleles and on recombination that selectively reduces that association as a function of the marker distance from the QTL. In traditional QTL mapping, usually called linkage mapping, the creation and selective removal of LD both occur within the boundaries of the experiment. Linkage disequilibrium is created by hybridization between inbred lines and decays through recombination during the production of recombinant progeny. In association mapping, both processes occur outside the boundaries of the experiment and are therefore not under experimental control. The primary mechanisms generating LD are mutation and drift, while recombination continues to be the sole systematic mechanism reducing LD.

The relatively few generations of recombination that are used in standard linkage mapping allow little opportunity for dissipation of LD. Consequently, high LD prevails even between distant loci (e.g., 10 cM, which would be very roughly 50×10^6 base pairs for the *Triticeae*), and linkage

J.-L. Jannink, USDA-ARS, R.W. Holley Center for Agriculture and Health, Cornell Univ., Ithaca, NY 14850; H. Iwata, National Agricultural Research Center, Tsukuba, Ibaraki 305-8666, Japan; S. Chao, USDA-ARS, Biosciences Research Lab., 1605 Albrecht Blvd., Fargo, ND 58105-5674; P.R. Bhat, Dep. of Botany and Plant Sciences, Univ. of California-Riverside, Riverside, CA 92521; P. Wenzl, Triticarte P/L, P.O. Box 7141, Yarralumla (Canberra), ACT 2600, Australia; G.J. Muehlbauer, Dep. of Agronomy and Plant Genetics, Univ. of Minnesota, St. Paul, MN 55108. Received 4 Sept. 2008. *Corresponding author (jeanluc.jannink@ars.usda.gov).

Abbreviations: CAP, coordinated agricultural project; DArT, diversity arrays technology; LD, linkage disequilibrium; MAF, minor allele frequency; OPA, oligonucleotide pool assay; OWB, Oregon Wolfe barley; QTL, quantitative trait loci; SNP, single nucleotide polymorphism.

mapping is inadequate to the task of fine mapping QTL. In contrast, in association mapping, many cycles of recombination may occur prior to the experiment such that LD may decay over a quite short span. For the detection of QTL to succeed, marker density must be matched to the rate of decay of LD. The key LD parameter to consider here is the r^2 between loci because the fraction of the phenotypic variance that a marker will explain is directly related to its r^2 with the QTL. In linkage mapping with biparental crosses, the relationship between r^2 and genetic distance is straightforward, in part because allele frequencies are always close to 0.5. For mapping within F_2 or doubled-haploid populations, $r^2 \approx 1 - 4c$, where c is the recombination frequency. This approximation holds reasonably well for $c < 0.1$. In association mapping, LD is not nearly as well behaved. In an equilibrium population at effective population size N_e, the *expectation* of r^2 is $E(r^2) = 1 / (4N_e c + 1)$ (Hill and Robertson, 1968). There will be quite a bit of variability around this expectation caused by the fact that allele frequencies at loci will differ from each other and because drift, by definition, creates LD in random ways. Given these caveats, assuming an effective population size for cultivated barley (*Hordeum vulgare* L.) of 100 would mean that average LD would decay to quite low levels ($r^2 = 0.20$) within 1 cM. The actual extent of LD, however, will need to be measured in each study population. In particular, the meaning of "effective population size" is unclear for a species in which the "population" consists of a mosaic of breeding programs. Nevertheless, as a rough guide, this calculation shows that population-wide LD can decay much more rapidly than the family-based LD of standard QTL mapping populations. Required marker density will therefore also be higher.

To develop computational strategies to alleviate marker density needs in plants, it is useful to look to human genetics where marker densities are much greater and tremendous resources have been invested. Given that human is an outbred diploid, a first issue is to identify marker phase (that is, for a series of heterozygous markers, to identify which alleles have the same parental origin). Methods to infer the phase of markers have also been used to impute missing marker data (Scheet and Stephens, 2006). The process of imputing has been extended to impute the allelic state of markers that were not typed in the experimental dataset (Servin and Stephens, 2007; Marchini et al., 2007). This imputation requires a reference panel that has been densely typed to provide a sample of haplotypes that include markers not typed in the experimental dataset (Fig. 1). The imputation algorithm implemented in the software fastPHASE (version 1.3) applies a statistical model to the data that clusters haplotypes identified over short stretches of the genome (Scheet and Stephens, 2006). When marker data is missing, haplotypes are assigned to clusters using available data, and missing scores are imputed to the allele most frequent for that cluster (Scheet and Stephens, 2006). Missing marker scores may occur in random cells of the genotype × marker matrix, as when a particular genotype/marker combination fails, or it may be systematic, as when all experimental data is not typed for a particular marker such that it is only available in the reference panel.

Superficially, the marker imputation described seems to have little to do with the process of tag single nucleotide polymorphism (SNP) selection, where certain SNP are chosen as proxies for others with which the proxies

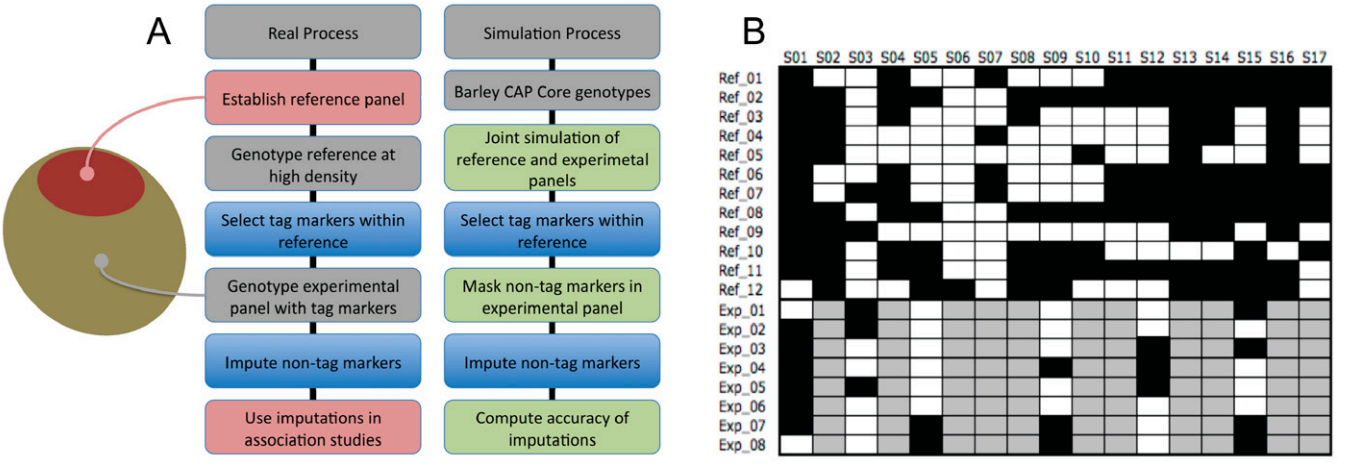

Figure 1. A. Left: Sequence of the real process through which imputation methods would be used in association studies, from the selection of a reference panel through the use of imputed marker scores in association studies. Right: Parallel view of the simulation process used to verify that this approach might be fruitful. The conceptual experimental population in which imputation would be performed is in beige with the reference panel selected from it in red. Key common steps between real and simulation processes are highlighted in blue. Simulation steps that allow verification of imputation accuracy are in green. Steps in the real process that would benefit from further research are in red. B. Illustration of a small portion of the structure of reference and experimental panels for the imputation of markers that have not been scored on experimental lines. Markers (S01 to S17) are in columns and barley lines in rows. Markers in gray have not been scored in the experimental panel (denoted by line names "Exp_##") and will be imputed on the basis of information in the reference panel (denoted by line names "Ref_##") and on tag marker (S01, 03, 05, 09, 12, and 15) data in the experimental panel.

are in high LD (Stram, 2004; de Bakker et al., 2005). But in fact, tag SNPs are also used to impute the allelic state of non-tag markers. This method of imputation is simpler than that of the clustering algorithm of fastPHASE. Specifically, the allele of a non-tag SNP is simply taken to be in agreement with that of its tag SNP proxy. Multiple tag SNP combinations are also possible (de Bakker et al., 2005) and they also generate imputations by simple one-to-one correspondences between a tag SNP combination and a non-tag SNP allelic state. Ultimately the outcome is the same for both tag SNP and clustering approaches: some markers are typed while others are not. Non-typed markers are imputed, and the association analysis regresses the phenotype on the allelic states of all markers, typed or imputed.

High density DNA marker data are becoming available in plants and a few genome-wide association mapping studies have been published (eg., Kraakman et al., 2004, 2006; Crossa et al., 2007; Steffenson et al., 2007). Thus, methodologies for imputing markers will be increasingly needed. To our knowledge, the clustering methodology of fastPHASE has not been applied to impute markers in a crop before. Given the non-natural population structure characteristic of crops (Hamblin et al. 2005; Hamblin et al., 2006; Rostoks et al., 2006), we wanted first to evaluate the effectiveness of fastPHASE at imputing missing marker scores when these scores were randomly missing from a dataset. We also compared, by simulation, imputation accuracy obtained from designed tag SNP proxy tests to that of fastPHASE when certain non-tag markers were systematically missing from the experimental dataset (Fig. 1A). Given the possibility of using fastPHASE to replace tag SNP tests, we evaluated two methods of choosing tag SNP (de Bakker et al., 2005) for their ability to allow fastPHASE to impute accurately. Finally, since most current SNP sets in crops will not have been explicitly chosen as tag SNP, we looked at two methods of supplementing a random set of SNP markers with few additional markers so as to maximize fastPHASE imputation accuracy. To tie these evaluations to the complexity of crop LD patterns, all analyses were based on marker data generated on barley by the Barley Coordinated Agricultural Project (www.barleycap.org).

Materials and Methods

Germplasm and Marker Data

The data analyzed in this study were derived from marker polymorphisms on the barley CAP core. The CAP core consists of a set of 102 lines containing primarily North American lines (there are 11 lines of non-North American origin). Barley can be grouped according to row type (2- or 6-row), growth habit (spring or winter), and usage (forage, feed, or malt). Grouped in this way, the CAP core contains 30 2-row and 27 6-row spring malt types; one 2-row and two 6-row winter malt types; 11 2-row and three 6-row spring feed types; 11 6-row winter feed types, four forage types, and 13 genetic stocks. These CAP core lines have been screened with SNP obtained from three Illumina GoldenGate oligonucleotide pool assays (OPA). One assay was described in Rostoks et al. (2006), denoted pilot OPA1 (POPA1), and the two others were developed with similar methods (T.J. Close, personal communication, 2008), denoted POPA2 and POPA3. Ninety-five of the lines were also screened with barley Diversity Arrays Technology (DArT) markers (Wenzl et al., 2004). In preliminary analyses using these markers, four lines ('Oregon Wolfe Barley-Dominant,' 'Oregon Wolfe Barley-Recessive,' 'Steptoe,' and 'Barke') did not cluster well with the others and were excluded from further analysis. Of the remaining 98 lines (Table 1), 'Haruna Nijo,' 'BCD12,' and 'Morex' were not scored with the DArT markers. After removing markers for which more than half the lines had missing scores, there were 1399, 1253, 1216, and 1476 POPA1, POPA2, POPA3, and DArT markers, respectively. Of these, respectively 1030, 930, 769, and 1100 had map positions from consensus bi-parental mapping projects (T.J. Close personal communication, 2008; Wenzl et al. 2006, P. Szucs and P. Hayes, personal communication, 2008). Sequence and marker information were used to remove redundant DArT from the dataset. Clones from 1265 DArT markers were sequenced. All markers whose clones clustered on the basis of their sequence were considered redundant. In two cases, DArT were joined to a cluster by their sequence but their marker scores differed from the cluster consensus for more than five lines. Those DArT were considered non-redundant. For unsequenced DArT markers, if their scores across the barley core matched exactly, they were also considered redundant. For all redundancy groups, the single marker with the fewest missing scores was retained. This process left 3131 markers with map positions. The POPA map containing positions of 2943 EST-derived SNP was used as the reference map (T.J. Close, personal communication, 2008; and Harvest:Barley, http://harvest.ucr.edu and http://www.harvest-web.org). To merge DArT markers onto this map, the DArT consensus map (Wenzl et al., 2006) and the 2383-locus Oregon Wolfe Barley (OWB) map (P. Szucs and P. Hayes, personal communication, 2008; http://www.barleycap.org), containing both POPA and DArT markers, were used. The following expedient approach was used to merge these maps. For each chromosome, common markers between the OWB and DArT maps were identified. Per chromosome, there were on average 69 (range: 32–94) common markers. For common markers, OWB positions were regressed on DArT positions. For remaining markers, that regression was used to project DArT positions onto the OWB map (OWB map positions remained unchanged in this merge). The same procedure was used to project OWB and DArT positions onto the reference POPA map. Finally, for 102 unmapped POPA SNP whose scores matched perfectly with those of a mapped marker, the unmapped marker was also placed at its matching marker's position. The rate of missing

Table 1. Names of lines in the barley CAP core, the row subpopulation they were assigned to according to their spike phenotype (two or six row), and the subpopulation label given to them for the imputation of missing marker scores (see Materials and Methods). Lines that were not assigned a row subpopulation were not used in analyses involving these subpopulations.

No.	Line	Row subpop.	Label	No.	Line	Row subpop.	Label
1	2B96-5038	two	1	50	FEG66-08	six	3
2	2B98-5312	two	2	51	FEG90-31	six	3
3	6B00-1526	six	3	52	Flagship	two	2
4	6B02-3394	six	3	53	Foster	six	3
5	6B94-7378	six	3	54	Franklin	two	2
6	6B94-8253	six	3	55	Garnett	two	1
7	6B97-2245	six	3	56	Geraldine	--	4
8	88Ab536	six	3	57	Harrington	two	1
9	88Ab536-B	six	3	58	Haruna Nijo	two	2
10	AC Metcalfe	two	1	59	Haxby	two	1
11	Arapiles	two	2	60	Hays	--	4
12	B1202	two	2	61	Hockett	two	1
13	B1215	two	2	62	Hoody	--	6
14	B1602	six	3	63	Klages	two	1
15	B1614	six	3	64	Kold	--	6
16	Baronesse	--	4	65	Kompolti	--	6
17	BCD12	two	1	66	Lacey	six	3
18	BCD47	two	1	67	Larker	six	3
19	Belford	--	5	68	Legacy	six	3
20	Bison 1H	--	4	69	M122	six	3
21	Bison 1H+4H	--	4	70	M123	six	3
22	Bison 1H+4H+5H	--	4	71	Merit	two	1
23	Bison 1H+5H	--	4	72	MNBrite	six	3
24	Bison 4H	--	4	73	Morex	six	3
25	Bison 4H+5H	--	4	74	ND20448	six	3
26	Bison 5H	--	4	75	ND20508	six	3
27	Bison 7H	--	4	76	ND21863	two	1
28	Bowman	two	1	77	NDB112	six	3
29	C-14	two	2	78	Newdale	two	1
30	Canela	--	2	79	Nomini	--	6
31	CDC Copeland	two	1	80	Orca	two	1
32	CDC Kendall	two	1	81	Pasadena	two	2
33	CDC Sisler	six	3	82	Price	--	6
34	CDC Stratus	two	1	83	Radiant	--	4
35	Charles	two	1	84	Rawson (ND19119-2)	two	1
36	CIho 4196	two	2	85	Robust	six	3
37	Collins	two	2	86	Scarlett	two	2
38	Conlon	two	1	87	Shenmai 3	--	2
39	Conrad	two	2	88	Stander	six	3
40	Craft	two	1	89	Stellar	six	3
41	Crest	two	2	90	Strider	--	6
42	Dicktoo	--	6	91	Sublette	two	1
43	Doyce	--	6	92	Sussex	--	6
44	Drummond	six	3	93	Thoroughbred	--	6
45	Eslick	--	4	94	TR306	two	1
46	Excel	six	3	95	Tradition	six	3
47	Farmington	two	2	96	WA1614-95	--	6
48	FEG55-14	six	3	97	Washford	--	5
49	FEG59-09	six	3	98	Wysor	--	6

marker data was 1.8% overall (3.2% for the DArT markers and 0.5% for the SNP markers).

Marker Imputation

Missing marker data were imputed using the program fastPHASE version 1.3 (Scheet and Stephens, 2006). Briefly, the fastPHASE algorithm works as follows: Haplotypes in the population are assumed to cluster, with each locus having characteristic allele frequencies within each cluster. The observed haplotypes are assumed to arise as a mosaic of segments originating from different clusters. Alleles at adjacent loci usually originate from the same cluster, but there is a transition probability of shifting origin from one cluster to another between each locus. The likelihood of the model is maximized for the cluster probability, allele frequency, and transition probability parameters to obtain maximum likelihood parameter estimates. If a marker data point is missing, the probability of it being one allele or the other is calculated as a function of the cluster of origin probabilities of the haplotype at that marker and the allele frequencies for the marker in each cluster. The most likely allele is the one imputed. If the lines in the dataset are known to come from different subpopulations, different cluster probabilities and transition probabilities can be estimated for each subpopulation (Scheet and Stephens, 2006).

To perform an analysis, 2% of the marker data was masked (the score was marked as missing). The program fastPHASE was then used to impute missing data, with options to indicate that all haplotype phases were known (since the data were from inbred lines), and that subpopulation labels were used. We also attempted imputation allowing for a single pseudo-recombination parameter, but this gave poor accuracies (data not shown). Haplotype cluster numbers of 10, 20, 30, and 40 were tested. The fastPHASE imputed marker score was compared to the known score of the masked data point and each marker was characterized according to the frequency with which its missing scores were correctly imputed. This operation was done 200 times for each of the seven barley chromosomes. We first compared the program Structure (Pritchard et al. 2000) to K-means clustering (Hartigan and Wong, 1979) to assign lines to subpopulations, assuming K = 4 subpopulations. Structure (Pritchard et al., 2000) was run using the no admixture model. Imputation accuracies when lines were assigned

to subpopulations by Structure versus K-means clustering were virtually identical, differing on average across haplotype cluster numbers by only 0.02%. Because line assignment was more stable by K-means clustering, we proceeded with it. Subpopulation numbers of 1, 2, 4, 6, and 8 were tested in factorial with haplotype cluster numbers of 10, 20, 30, and 40.

Comparison of fastPHASE with Designed Tag SNP Tests

The raw marker data used here are unlike what might be encountered in an actual breeding program in the sense that the germplasm is of more than one type and of somewhat diverse geographical origins. To generate datasets of the size and type that a breeder might actually work with, we simulated one generation of random mating within the largest two-row and six-row subsets of the base population that were identified using K-means clustering assuming four subpopulations (Table 1, Fig. 2). Progeny were simulated by randomly pairing base population genotypes into an F_1 and generating a gamete following Mendelian segregation and rules of recombination assuming independent crossovers. The gametes represented possible genotypes of the progeny generation. Five simulations were performed on each subset. The analyses distinguished between a reference panel and an experimental panel (Fig. 1B). The assumption is that, in the real world, all marker data are available on the reference panel, but the experimental panel will only be typed for the tag SNP. In our case, through simulation, we also had scores for non-tag SNP available on the experimental panel (Fig. 1A). Reference panel sizes of 100 and 200 inbreds were simulated. Experimental panels were always of 200 inbreds.

Working one chromosome at a time, reference panel marker data were submitted to Haploview (Barrett et al., 2005) for tag SNP selection. Haploview provides two tag SNP identification methods. In the greedy method (Barrett et al., 2005), a marker is ranked according to the number of other markers with which it is in LD at an r^2 above a minimal specified value that we called the tag selection r^2. The highest ranked marker is picked as a tag. The remaining markers are then re-ranked according to the number of non-tag markers they predict and the best is added to the tag set. This process is iterated until all SNP not included in the tag set (non-tag SNP) have an r^2 of at least a specified level with a tag SNP. Aggressive tagging can be accomplished, allowing non-tag SNP also to be predicted by a combination of two or three tag SNP. We employed this aggressive up-to-three SNP tagging and specified a minimal r^2 level of 0.4. In the bestN method (Barrett et al., 2005), a maximum number of tag SNP is specified, as well as a minimal r^2. Tag SNP are ranked and picked in the same way, but only up to the maximum number of tags allowed (that is, not all non-tag SNP will be predicted at a minimal level). For the bestN method, we chose the maximum number of tags as the number of tag SNP identified by the greedy method, but used a minimal r^2 of 0.8.

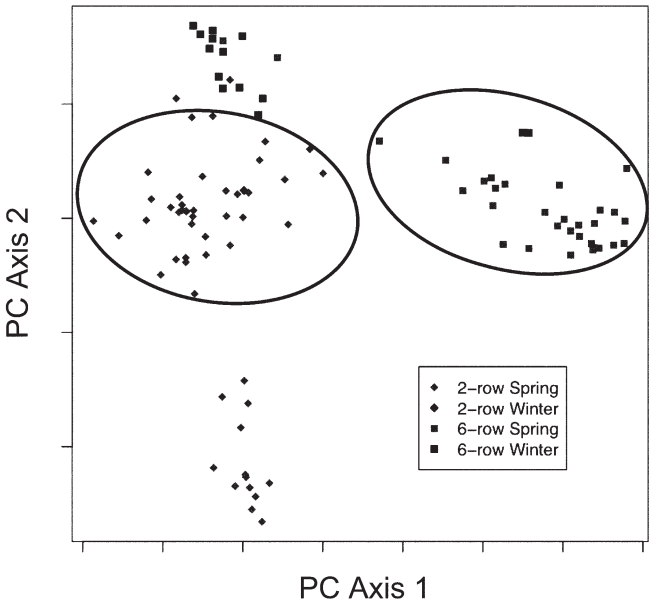

Figure 2. Lines of the barley CAP core plotted according to the first two eigenvectors determined by principal component analysis. The left and right ovals surround, respectively, the 2- and 6-row subpopulations used as parents in simulating populations for evaluation of imputation accuracy.

In picking the tags, Haploview also designs proxy tests that predict the score of each non-tag SNP based on the allelic states of single tag SNP or specific combinations of two or three tag SNP (Barrett et al., 2005). To determine the accuracy of Haploview predictions, the r^2 between the proxy test prediction and the true marker score for non-tag SNP was calculated on the experimental panel that had not been part of tag identification. Note that this r^2, which we call the non-tag prediction r^2, is different from the tag selection r^2 previously described. The tag selection r^2 is calculated in the reference dataset while the non-tag prediction r^2 is calculated in the experimental dataset. The non-tag prediction r^2 can only be calculated because of the simulation setup that we used. To determine the accuracy of fastPHASE imputations, a dataset concatenating the reference and experimental panels was analyzed. For each non-tag SNP, the prediction r^2 was calculated between fastPHASE imputations and masked scores in the experimental panel.

Finally, we sought to combine information from Haploview proxy test predictions with fastPHASE predictions as follows. First, the probability that the marker score b was 1 was calculated conditional on the proxy test prediction a using

$$P(b=1|a=0)=P_b-r_t\sqrt{\frac{1-P_b}{1-P_a}P_aP_b}$$

and

$$P(b=1|a=1)=P_b+r_t\sqrt{\frac{P_b}{P_a}(1-P_a)(1-P_b)}$$

where P_b is the frequency of the 1 allele at the non-tag SNP in the reference panel, P_a is the frequency of the 1 allele prediction in the experimental panel, and r_t is the square root of the coefficient of determination of the

proxy test in the reference panel. These equations were derived from

$$r_e = \frac{P_{ab} - P_a P_b}{\sqrt{P_a(1-P_a)P_b(1-P_b)}} \text{ and } P(b=1 \mid a=1) = \frac{P_{ab}}{P_a}$$

or

$$r_e = -\frac{P_{\bar{a}b} - (1-P_a)P_b}{\sqrt{P_a(1-P_a)P_b(1-P_b)}} \text{ and } P(b=1 \mid a=0) = \frac{P_{\bar{a}b}}{1-P_a}$$

where r_e is the (unobserved) square root of the coefficient of determination of the proxy test in the experimental population and we estimate r_e by r_t; P_{ab} and $P_{\bar{a}b}$ are, respectively, the joint probabilities of $a = 1$ and $b = 1$ and $a = 0$ and $b = 1$. Second, the probability that marker score b was 1 on the basis of fastPHASE analysis was calculated from the frequency with which fastPHASE returned 1 over 100 calls. A weighted average of these two probabilities was calculated. The weight varied for each non-tag marker and depended on three independent variables: the marker's minor allele frequency, and the coefficients of determination of the proxy test and of fastPHASE predictions in the reference panel. The probability that Haploview was correct was calculated as the number of times that was true divided by the total number of disagreements. This variable was regressed on the three independent variables. The weight given to the Haploview marker score imputation was taken as its regression-based predicted probability of being correct. If the weighted average was greater than 0.5, a 1 allele was imputed and otherwise a 0 allele was imputed.

Imputation of Missing Markers on the Basis of Random Tag SNP

The ability to select tag SNP is predicated on the fact that, initially, a reference panel is available that is scored at very high density. An optimal subset of the markers on that panel is then selected as tag SNP and these tags are scored on experimental data. Given the current development of SNP in barley, we are more likely to have a reverse situation: about 3000 markers that have not been specifically selected as tags will be scored on a large experimental population. Only subsequently might we be able to develop additional markers, score a reference panel at even higher density, and then seek to impute the new markers on the experimental population, using the 3000 currently existing SNP as a guide. This reverse situation is like having random SNP as tags. We therefore wanted to assess the accuracy of marker imputation when available data were not from carefully selected tag SNP but from random SNP. Still working one chromosome at a time, datasets for imputation by fastPHASE were created by concatenating to the reference panel the marker scores of randomly picked SNP in the experimental panel. The prediction r^2 between fastPHASE imputations and masked experimental panel SNP were then calculated. The number of SNP retained at random was 20% of the total number of polymorphic SNP. In a second approach, the SNP retained were not a random set and instead only every 20th SNP was dropped and left to be imputed. Thus, 95% of SNP were retained and those SNP dropped were well-surrounded by scored markers.

In order to improve imputation of new markers, we assumed it would be possible to score the experimental dataset with a smaller series of markers selected using the high-density panel. We compared two methods of selecting the extra "imputation support" markers. First, we used Haploview to pick extra tags. Again working one chromosome at a time, Haploview was used with an option forcing it to include the initial random SNP among the tag SNP, but to add either ten or twenty markers to the random set. Thus the random SNP would be supplemented by ten or twenty specifically selected tag SNP per chromosome, which, in barley, would amount to scoring an additional 70 or 140 markers on the experimental dataset. Second, for each SNP in the high-density panel, we determined how difficult it was to impute using fastPHASE by calculating the r^2 between its true and imputed scores. The ten or twenty SNP per chromosome that had the lowest r^2 were then taken as the imputation support markers. Having selected imputation support markers, fastPHASE imputation datasets were created with the reference panel and experimental panel typed only with random plus support SNP. Again, the r^2 between fastPHASE imputations and the true scores of SNP missing from the dataset were then calculated.

Results

Missing Marker Imputation Accuracy

The imputation accuracy of fastPHASE was quite high. The highest correctness, averaged over all markers, was 97.1% and occurred assuming a single subpopulation and thirty haplotype clusters (Fig. 3). The standard deviation of correctness across markers in that case was 3.4%. Nearly 80% of markers were imputed correctly more than 95% of the time, and nearly 50% were imputed correctly more than 98% of the time (Fig. 4). Assuming at least 20 haplotype clusters were modeled, however, correctness never went below 96.7%, showing that it was not highly sensitive to subpopulation and cluster number. We hypothesize that assuming more than one subpopulation only becomes useful if each subpopulation will have a sufficient number of individuals in it. Since we only had 98 individuals to start with in the barley CAP core, that threshold was not reached.

Tag Selection and Imputation Using Tag SNP

Differences between the two subpopulations (two- and six-row) caused tag SNP selection to differ between them (Table 2). The two-row subpopulation was more polymorphic than the six-row. On average over simulations there were 2495 and 1595 markers with minor allele frequency > 0.05 for the two- and six-row subpopulations, respectively. Linkage disequilibrium was also less in the two- than the six-row subpopulation: a higher percentage of markers chosen as tags was required to predict non-tag

markers at the minimum tag selection r^2 within the two- than the six-row subpopulation (Table 2). The difference between the subpopulations in the percentage markers needed as tags increased as the minimum tag selection r^2 increased from 0.4 to 0.8. As a result of these two facts, the number of tag SNP needed for the two-row subpopulation was more than double that needed for the six-row subpopulation. For example, for the tag selection r^2 of 0.4, 610 and 273 markers, on average, were needed for the two- and six-row subpopulations. Once tag SNPs were selected, however, the two subpopulations behaved quite similarly: non-tag markers were predicted with similar r^2 (Table 3). There were, however, very large differences between the methods in their ability to predict scores for non-tag markers (Fig. 5A; Table 3). In particular, imputations from fastPHASE matched the true values of marker scores with much higher r^2 than did the proxy test predictions designed by Haploview. The minimal prediction r^2 were similar for all methods, with more than 95% of all markers being predicted at an r^2 better than 0.3 (Fig. 5A). From there, however, the methods diverged with fastPHASE predicting much higher percentages of markers at high r^2 than the Haploview-designed proxy tests (Fig. 5A, Table 3). In fact, in general, fastPHASE did a better job of predicting non-tag markers even when it used only randomly-selected tag SNP, than Haploview did with its carefully chosen tags. Besides this major differentiation between the methods, there was also an interaction between method and subpopulation: Haploview tended to do a better job of prediction in the six-row than the two-row subpopulation while the reverse was true for fastPHASE (Table 3). This interaction was, however, swamped by the overall superiority of fastPHASE imputations.

Over all marker scores where the Haploview proxy test and fastPHASE disagreed, the proxy test was correct only 7% of the time in both two- and six-row subpopulations. Three variables were significantly correlated to the probability that the proxy test was correct: the marker's minor allele frequency (MAF) and the coefficients of determination of the proxy test and of fastPHASE predictions in the reference panel. Multiple regression showed that in case of disagreement, decreasing MAF and increasing proxy test coefficient of determination increased the probability that Haploview was correct, while increasing fastPHASE coefficient of determination decreased that probability. For the two- and six-row subpopulations however, only 14 and 8%, respectively, of the variation in that probability was explained by these variables (Fig. 6). The regression-predicted probability of Haploview correctness never exceeded 50% (Fig. 6). Consequently, our attempts to use Haploview predictions to increase correctness of fastPHASE imputations in cases where the two disagreed were unsuccessful: the raw fastPHASE prediction was always better than a prediction from the combination of Haploview and fastPHASE predictions (data not shown).

Tag Selection Methods Appropriate for fastPHASE Imputation

The previous results show that tag SNP selected using Haploview's greedy algorithm provide an excellent basis for fastPHASE's imputation algorithm. Looking at the power of QTL detection, de Bakker et al. (2005) found that their bestN algorithm provided higher power than the greedy algorithm. We tested whether either algorithm provided tags leading to higher imputation accuracy (Fig. 5B). There were no important interactions between the tag selection algorithm and the

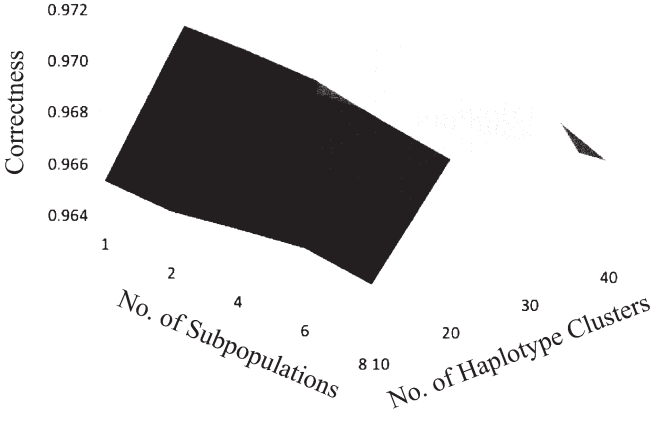

Figure 3. Fraction of masked marker scores correctly imputed by fastPHASE as a function of the number of subpopulations assumed in the barley CAP core and the number of haplotype clusters modeled in the analysis. Note the very small range of less than 1% between the best and worst correct fractions.

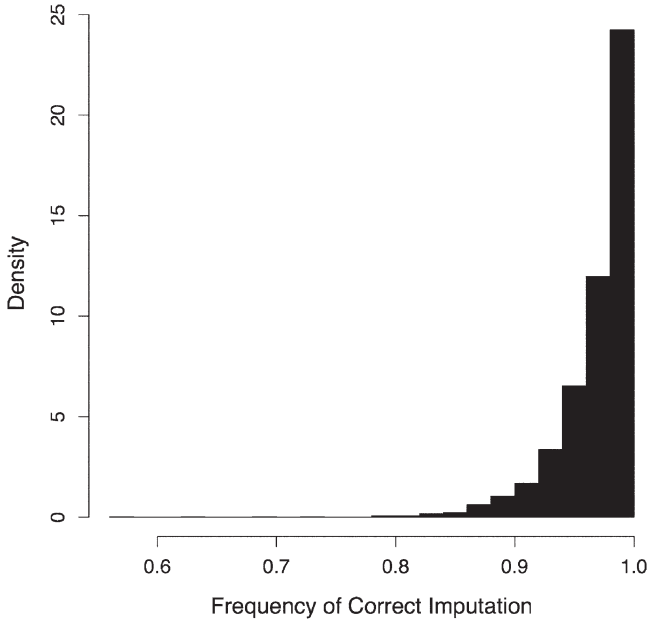

Figure 4. Probability density function of the distribution across markers of imputation correctness when marker scores missing at random were imputed by fastPHASE. Each point in the distribution represents for a marker the frequency that the marker's scores were correctly imputed.

Table 2. ANOVA and mean percentages of markers selected as tags using Haploview's greedy method for two- and six-row subpopulations, reference panel sizes of 100 or 200 inbreds, and minimum tag selection r^2 of 0.4, 0.6, and 0.8.

		Percentage markers used as tags		
Source	df	Mean Square		
Subpopulation (S)	1	0.202 ***		
Ref. Panel Size (P)	1	0.002 ***		
Min. Tag. Sel. r^2 (R)	2	0.210 ***		
S x P	1	0.000		
S x R	2	0.002 ***		
P x R	2	0.000		
S x P x R	2	0.000		
		Means by subpopulation and minimum tag selection r^2 [%]		
		$r^2 = 0.4$	$r^2 = 0.6$	$r^2 = 0.8$
Two-row		24.5	34.6	45.9
Six-row		17.1	23.6	32.7

***Significant at the 0.001 probability level.

subpopulation or the reference panel size, so we present means across all analyses (Fig. 5B). First, to determine the upper limit of imputation accuracy at the available marker density, we used fastPHASE to impute markers when very few markers (one out of 20) were masked in the dataset. In that case, 93% of markers are predicted at an r^2 of 0.8 or greater (Fig. 5B). The greedy algorithm did better than the bestN algorithm at ensuring that all non-tag markers were imputed above a minimal r^2. For example, when using the greedy algorithm, over 96% of markers were imputed at a prediction r^2 of 0.6 or better, whereas this percentage was just over 91% for the bestN algorithm (Fig. 5B). In contrast, the bestN algorithm predicted more markers than the greedy algorithm at a very high r^2 level: 68 versus 60% of markers were predicted at an r^2 of 0.9 or higher using the bestN and greedy algorithms, respectively.

Table 3. ANOVA and mean percentages of markers imputed with a prediction r^2 greater than 0.5 and 0.8. Analyses were fastPHASE imputations on tag SNP (fP/tag) or on randomly selected SNP (fP/rand), and Haploview proxy test predictions on tag SNP (HV/tag).

		Percentage markers with					
		Prediction $r^2 \geq 0.5$			Prediction $r^2 \geq 0.8$		
Source	df	Mean square			Mean square		
Subpopulation (S)	1	0.001			0.038 *		
Ref. Panel Size (P)	1	0.085 ***			0.114 ***		
Analysis (A)	2	5.104 ***			7.701 ***		
S x P	1	0.007			0.000		
S x A	2	0.238 ***			0.172 ***		
P x A	2	0.004			0.035 **		
S x P x A	2	0.017			0.008		
		Means by subpopulation and analysis [%]					
		fP/tag	fP/rand	HV/tag	fP/tag	fP/rand	HV/tag
Two-row		98.5	90.3	71.7	81.1	65.9	31.2
Six-row		97.8	85.6	79.3	77.9	66.0	40.4

***Significant at the 0.001 probability level.

Optimally Supplementing Random Tags

If researchers do not have the luxury of careful selection of tag SNP at the outset, it may still be possible to supplement marker scores on a previously genotyped experimental dataset with a small number of additional markers to improve imputation of many more markers from a newly-established reference panel. Adding only ten markers per chromosome was effective at improving the ability of fastPHASE to impute remaining markers. Adding ten extra markers that were tough to impute or that were selected by Haploview's bestN method increased the percentage of markers predicted at an r^2 of 0.8 or better by 8 and 12%, respectively (Fig. 5C). When adding twenty extra markers the increases for the two methods were of 15 and 19%, respectively. Particularly for maximizing the number of markers imputed at high r^2 levels, the Haploview-selected markers were more effective tags than markers that fastPHASE imputed poorly (Fig. 5C).

Discussion

The result that missing scores for about half of the markers in our dataset were imputed correctly 98% of the time or better indicates that imputation using local clustering as implemented in fastPHASE (Scheet and Stephens, 2006) can be a useful tool for dealing with missing marker data. We believe that these imputation accuracies are conservative. First, the barley CAP core does not represent a "population" in the way that lines from an actual breeding program would. The CAP core therefore does not meet fastPHASE assumptions as well. For example, marker scores on winter barley lines in the core were imputed correctly only 93.1% of the time as compared to 97.6% of the time for spring lines. This difference was presumably simply caused by the fact that there are fewer winter lines in the core. Lines from a single program would not have this level of heterogeneity. Guan and Stephens (2008) also found that poorly-matched reference and experimental panels decrease imputation accuracy. Second, fastPHASE assumes that the correct marker order is known, whereas in our case marker order was approximate. Finally, the CAP core dataset used here represents only 98 lines, a relatively small reference panel.

The fastPHASE algorithm was clearly better at predicting non-tag marker scores than was the proxy test designed by Haploview (Fig. 5A). The resulting recommendation for tagging methods appears ironic: use Haploview to select tag SNP on the reference panel, but then disregard its own tests for applying tag SNP information and instead rely on fastPHASE to predict non-tag marker scores in an experimental panel. FastPHASE imputations had a further advantage over Haploview proxy tests in that the latter could not provide a prediction for

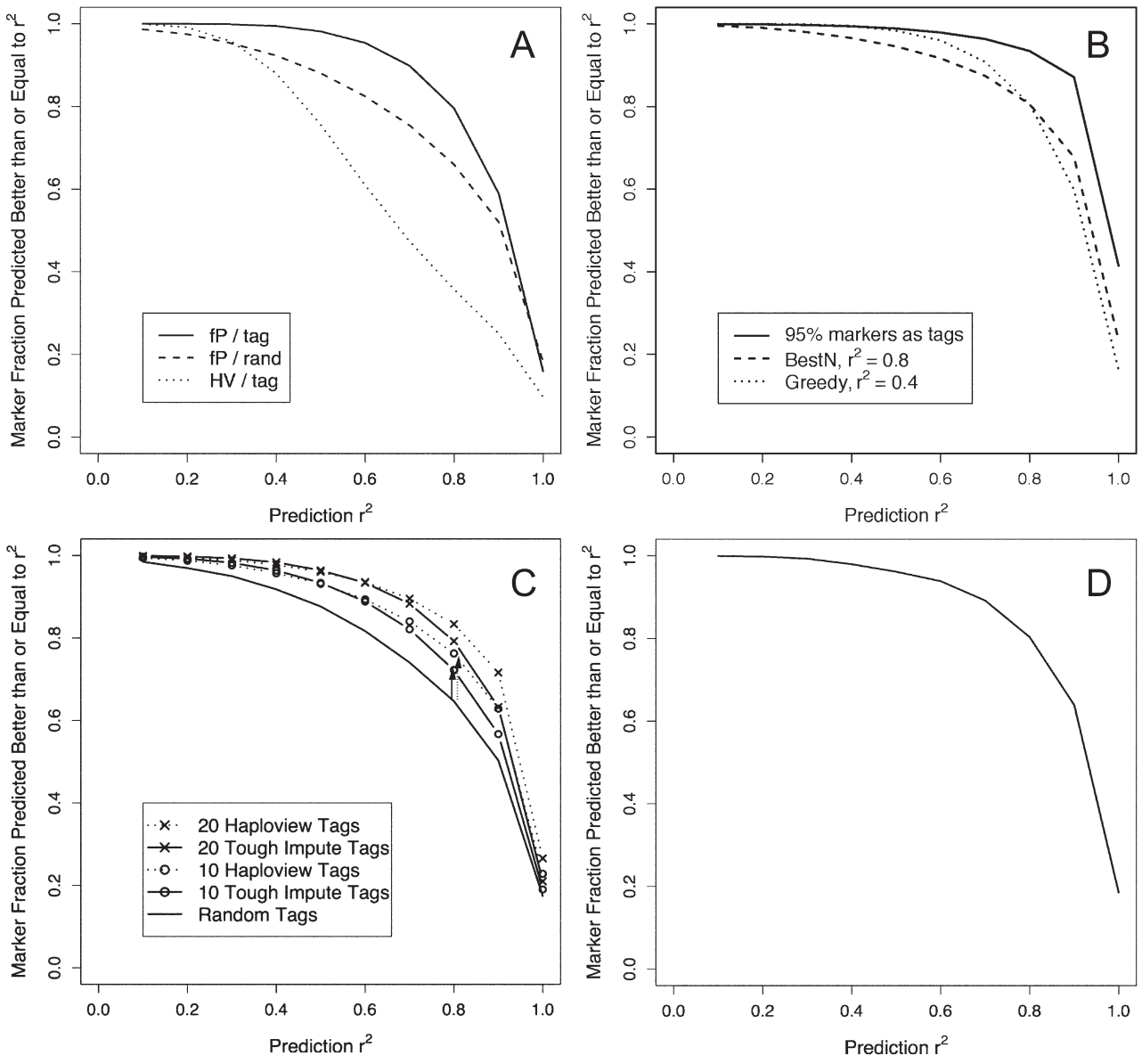

Figure 5. A. Fraction of non-tag markers predicted better than or equal to a range of prediction r^2. Prediction r^2 calculated for: Solid line – fastPHASE imputing markers based on Haploview-selected tag SNP; Dashed line – fastPHASE imputing markers based on randomly-selected tag SNP; Dotted line – Haploview-determined proxy tests. B. Fraction of non-tag markers predicted better than or equal to a range of prediction r^2. Prediction r^2 calculated for: Solid line – 95% of markers retained as tags ; Dotted line – greedy algorithm (de Bakker et al., 2005) used with tag selection r^2 set to 0.4; Dashed line – bestN algorithm (de Bakker et al., 2005) used with tag selection r^2 set to 0.8. C. Fraction of non-tag markers predicted better than or equal to a range of prediction r^2. Prediction r^2 calculated for: Solid line, no symbols – 20% of markers randomly selected as tags; Circles – ten supplemental markers per chromosome; Crosses – twenty supplemental markers per chromosome; Solid lines with symbols – supplemental markers chosen because fastPHASE predicted them poorly; Dotted lines with symbols – supplemental markers chosen by bestN algorithm (de Bakker et al., 2005) with tag selection r^2 set to 0.8. Corresponding solid and dotted vertical arrows indicate the increase in the fraction of markers predicted better than or equal to a prediction r^2 of 0.8 when ten markers were added per chromosome. D. Fraction of non-tag markers predicted better than or equal to a range of prediction r^2. Prediction r^2 calculated for all OPA markers when DArT in the dataset were used as tag markers.

lines where the relevant tag marker scores were missing, whereas the former always gave an imputation. Moreover, the prediction r^2 of imputation were very good, even when as few as 20% of markers were selected as tags (Table 3). These high r^2 suggest that, at least for 6-row spring barley in North America, it should be possible to design a panel of between 300 and 600 SNP that, when coupled to imputation on the basis of an appropriate reference panel, should be able to capture a very large fraction of known genomic variation. Further improvement in imputation accuracy over what we observed might be obtained by a two-step process in which haplotype cluster analysis is performed only on the reference panel, and clustering results are subsequently applied to experimental panels (Guan and Stephens, 2008). While the CAP core might serve as a reference panel, it was

not specifically designed for this purpose and further research should explore the optimal design of a panel for this function. An important distinction has been made, however, between coverage of *known* variation versus coverage of *complete* variation (Bhangale et al. 2008). In the case presented here, all SNP and DArT scores represent known variation, which appears to be very well imputed, whereas a large resequencing project would be necessary to assess coverage of complete, but currently unknown, variation.

The fastPHASE algorithm was optimally effective when the experimental dataset was scored with carefully-selected tag SNP, but it also performed reasonably well when only data on randomly-selected markers was available (Fig. 5A). This observation implies that imputation might still be fruitfully applied to experimental datasets scored with un-selected markers, as long as the reference panel also carries those markers. To test this idea, we simulated experimental datasets scored only with DArT markers and asked how well SNP data could be imputed in such datasets on the basis of the CAP core reference panel. We found that over 80% of SNP could be imputed with a prediction r^2 of 0.8 or greater (Fig. 5D). The relative success of imputation here argues that one could use the reference panel as a kind of Rosetta Stone allowing information from one marker system to be converted to that of another (Servin and Stephens, 2007). This process might allow for valuable meta-analyses across datasets that would seem otherwise incompatible.

At this time, we have no better proposal for the selection of tag SNP than the tagger algorithms implemented in Haploview (Barrett et al., 2005; de Bakker et al., 2005). These tag selection methods were not developed with the fastPHASE algorithm in mind, and so they may be sub-optimal. In particular, the tagger algorithms operate primarily on the basis of pairwise marker relationships whereas fastPHASE defines clusters on the basis of multi-locus information (Scheet and Stephens, 2006). At the same time, obviously, tag selection ensures that markers are scored that carry information about most or all markers that were not scored. Consequently, it makes sense that these methods should work reasonably well. Furthermore, it was striking how faithfully characteristics of pairwise tag selection methods were reflected in the prediction curves produced by fastPHASE. In particular, the greedy algorithm that ensures that all markers are tagged above a minimal threshold generated imputations that were above a similar threshold for a very high fraction of markers (Fig. 5B). In contrast, the bestN algorithm that ensures that a maximal number of markers are tagged at a high level generated a higher number of accurately imputed markers (Fig. 5B). Nevertheless, given how valuable accurate imputation can be, further research into optimal tag selection for use by fastPHASE might be fruitful.

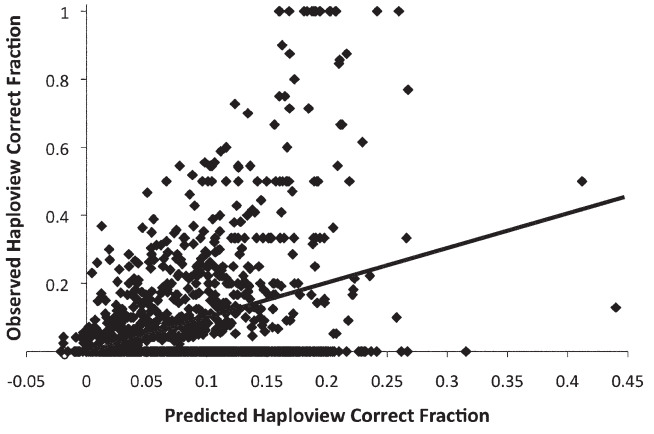

Figure 6. When Haploview and fastPHASE disagreed in their marker score imputation, the fraction of correct Haploview imputations was predicted by multiple regression (see Materials and Methods). Each point in the graph represents one marker. The line is the linear regression of observed on predicted Haploview correct fraction.

Finally, relative to variation that will be discovered in the future, our current markers and tags represent, in the worst case scenario, randomly selected tags (Bhangale et al. 2008). Consequently, after the discovery of new variation, we will want to supplement marker scores in experimental datasets with tags that will help impute this new variation (Fig. 5). Observations here provide optimism: when we supplemented random tags with 10 markers per chromosome (70 markers in total or 2.8 and 4.4% of the polymorphic markers in the two- and six-row subpopulations, respectively), an additional 12% of markers were imputed at a prediction r^2 above 0.8 (Fig. 5C). Thus, judicious selection of tags should, in this situation also, provide high returns on the investment.

In the study reported here, we have not addressed how researchers might use imputed scores obtained by fastPHASE. We distinguish between cases that require high confidence in the imputed allelic state versus cases where a continuous probability of allelic state would suffice. The former case might occur in marker-assisted selection where the researcher wanted to know the specific allele at a marker and therefore which scores were accurately imputed (e.g., belong to the set for which imputations are correct more than, say, 98% of the time). One approach would be to perform a simulation study such as done here in which available marker scores are masked and imputed, giving some indication of the markers for which actual missing data will be well imputed. Another approach would be to run fastPHASE multiple times on the same data set and to only retain imputed values that were consistent across all runs. In the present case, we found that marker scores that were imputed consistently over ten or more fastPHASE runs were imputed correctly 98.7% of the time. Similarly, it would be possible to impute missing marker data with more than one of the several methods currently proposed in the literature (Roberts et al. 2007; Sun and Kardia, 2008). Higher accuracy was found when the same score was imputed by more than one method (Sun and Kardia, 2008). Because Haploview imputations are not model-based (they rely on simple linkage-disequilibrium measures between tag and non-tag markers), whereas

fastPHASE imputations rely on a model assuming recombination processes, they might capture different signals from the data. We therefore attempted to combine the two to increase overall accuracy. We were unable to obtain a combination with higher accuracy than fastPHASE alone, however, primarily because when Haploview disagreed with fastPHASE, it was correct only 7% of the time. This low percentage did not allow sufficient margin for improvement of fastPHASE.

For some fraction of the "incorrectly imputed" cases we observed here, our dataset may be at fault because it contains genotyping errors. Indeed, imputation methods have been proposed to identify incorrectly scored markers and to improve allele calling algorithms (Marchini et al., 2007; Scheet and Stephens, 2008). In support of this idea, we found that consistently-imputed DArT markers were correct 97.4% of the time, while consistently-imputed SNP markers were correct 99.1% of the time. This difference may reflect lower error rates for SNP markers than for DArT markers, though other explanations are possible. For example, it may be that the map-merging procedure we used, combined with the fact that there were fewer DArT than SNP markers, increased the error of DArT marker placement, making DArT more difficult to impute. Of the scores actually missing in our dataset, 78% of them were imputed consistently to the same score over 200 fastPHASE runs. If we assume that those imputations were correct, we can reduce our percentage missing data rate from 1.8 to 0.4%, a valuable gain.

For many analyses, however, the algorithms do not require the exact allelic state but can be applied when imputation provides a probability that the allelic state is either 0 or 1. For example, singular value decomposition of the marker data matrix or regression of the phenotype on marker score can both use such continuous probabilities. Indeed, improved effect estimates may be obtained when the uncertainty in the data on the allelic state is taken into account by assigning a probability of allelic state rather than an all-or-nothing imputation (Dai et al. 2006; Kraft and Stram, 2007; Mensah et al., 2007; Guan and Stephens, 2008). Such a probability assignment can be obtained from fastPHASE by calling it repeatedly and taking a simple average of the imputations obtained from each call, or by using the –Pm option in fastPHASE version 1.4 or later (P.A. Scheet, personal communication, 2009). The software BIMBAM (Servin and Stephens, 2007; Guan and Stephens, 2008) can also output imputation probabilities directly. BIMBAM is designed for the analysis of human data and assumes heterozygous individuals, though it could presumably be fooled into working with inbreds (Y. Guan, personal communication, 2009). BIMBAM has not been tested yet for inbreds to our knowledge. Guan and Stephens (2008) have investigated association analysis using imputation and found that it can increase detection power relative to using only typed markers. Furthermore, imputation probabilities (that is, the posterior mean of a 0 or 1 imputation) were found to work as well as doing a full analysis in which imputation and association analyses are repeated to obtain a distribution of association outcomes across uncertain imputations.

We have shown that a recently-developed marker score imputation method developed by human geneticists in the context of association studies (Scheet and Stephens, 2006) can also work for a self-pollinating crop, despite the great differences in demographic histories between these different types of species. Simulations using real barley genotypes suggested that imputation will work well to alleviate problems associated with missing marker data and to increase the informativeness of tag markers. In the latter case, we found that for six-row North American barley it may be possible to score as few as 300 SNP and nevertheless retain a high degree of information on the scores of other common polymorphisms. While we have no direct evidence on the accuracy of fastPHASE in crops other than barley, the fact that the algorithm works well in species as divergent as humans and barley suggests that it should also work well in other crops. For breeding programs, cost savings associated with reducing marker densities to this degree could allow for additional applications such as earlier-generation marker screening and performance prediction. On the strength of these results based on genotypes alone, imputation methods appear to be quite useful for association mapping and breeding. Further evaluation, including the analyses of phenotypic data, is warranted.

Acknowledgments

We thank Paul Scheet for help with fastPHASE version 1.3. Martha Hamblin and Peter Bradbury gave valuable comments and advice over the course of this research. Patrick Hayes and Peter Szucs mapped DArT and SNP markers on the OWB mapping population used in this research. Timothy J. Close, Stefano Lonardi, and Yonghui Wu developed the 2943 SNP map used in this research. Sequencing of DArT marker clones was performed by Triticarte P/L. Comments of two anonymous reviewers helped greatly improve the manuscript. This research was supported by USDA-CSREES-NRI Grant No. 2006-55606-16722 "Barley Coordinated Agricultural Project: Leveraging Genomics, Genetics, and Breeding for Gene Discovery and Barley Improvement."

References

de Bakker, P.I.W., R. Yelensky, I. Pe'er, S.B. Gabriel, M.J. Daly, and D. Altshuler. 2005. Efficiency and power in genetic association studies. Nat. Genet. 37:1217–1223.

Barrett, J.C., B. Fry, J. Maller, and M.J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265.

Bhangale, T.R., M.J. Rieder, and D.A. Nickerson. 2008. Estimating coverage and power for genetic association studies using near-complete variation data. Nat. Genet. 40:841–843.

Crossa, J., J. Burgueno, S. Dreisigacker, M. Vargas, S.A. Herrera-Foessel, M. Lillemo, R.P. Singh, R. Trethowan, M. Warburton, J. Franco, M. Reynolds, J.H. Crouch, and R. Ortiz. 2007. Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics 177:1889–1913.

Dai, J., I. Ruczinski, M. LeBlanc, and C. Kooperberg. 2006. Imputation methods to improve inference in SNP association studies. Genet. Epidemiol. 30:690–702.

Guan, Y., and M. Stephens. 2008. Practical issues in imputation-based association mapping. PLoS Genet. 4:e1000279.

Hamblin, M.T., M.G. Salas Fernandez, A.M. Casa, S.E. Mitchell, A.H. Paterson, and S. Kresovich. 2005. Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. Genetics 171:1247–1256.

Hamblin, M.T., A.M. Casa, H. Sun, S.C. Murray, A.H. Paterson, C.F. Aquadro, and S. Kresovich. 2006. Challenges of detecting directional selection after a bottleneck: Lessons from *Sorghum bicolor*. Genetics 173:953–964.

Hartigan, J.A., and M.A. Wong. 1979. A K-means clustering algorithm. Appl. Stat. 28:100–108.

Hill, W.G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226–231.

Kraakman, A.T.W., R.E. Niks, P.M.M.M. van den Berg, P. Stam, and F.A. Van Eeuwijk. 2004. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:435–446.

Kraakman, A.T.W., F. Martínez, B. Mussiraliev, F. van Eeuwijk, and R. Niks. 2006. Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars. Mol. Breed. 17:41–58.

Kraft, P., and D.O. Stram. 2007. Re: The use of inferred haplotypes in downstream analysis. Am. J. Hum. Genet. 81:863–865.

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39:906–913.

Mensah, F.K., M.S. Gilthorpe, C.F. Davies, L.J. Keen, P.J. Adamson, E. Roman, G.J. Morgan, J.L. Bidwell, and G.R. Law. 2007. Haplotype uncertainty in association studies. Genet. Epidemiol. 31:348–357.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Roberts, A., L. McMillan, W. Wang, J. Parker, I. Rusyn, and D. Threadgill. 2007. Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. Bioinformatics 23:i401–i407.

Rostoks, N., L. Ramsay, K. MacKenzie, L. Cardle, P.R. Bhat, M.L. Roose, J.T. Svensson, N. Stein, R.K. Varshney, D.F. Marshall, A. Graner, T.J. Close, and R. Waugh. 2006. Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc. Natl. Acad. Sci. USA 103:18656–18661.

Scheet, P., and M. Stephens. 2006. A fast and flexible statistical model for large-scale Population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78:629–644.

Scheet, P., and M. Stephens. 2008. Linkage disequilibrium-based quality control for large-scale genetic studies. PLoS Genet. 4:e1000147.

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

Steffenson, B.J., P. Olivera, J.K. Roy, Y. Jin, K.P. Smith, and G.J. Muehlbauer. 2007. A walk on the wild side: mining wild wheat and barley collections for rust resistance genes. Aust. J. Agric. Res. 58:532–544.

Stram, D.O. 2004. Tag SNP selection for association studies. Genet. Epidemiol. 27:365–374.

Sun, Y.V., and S.L.R. Kardia. 2008. Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. Eur. J. Hum. Genet. 16:487–495.

Wenzl, P., J. Carling, D. Kudrna, D. Jaccoud, E. Huttner, A. Kleinhofs, and A. Kilian. 2004. Diversity arrays technology (DArT) for whole-genome profiling of barley. Proc. Natl. Acad. Sci. USA 101:9915–9920.

Wenzl, P., H. Li, J. Carling, M. Zhou, H. Raman, E. Paul, P. Hearnden, C. Maier, L. Xia, V. Caig, J. Ovesna, M. Cakir, D. Poulsen, J. Wang, R. Raman, K. Smith, G. Muehlbauer, K. Chalmers, A. Kleinhofs, E. Huttner, and A. Kilian. 2006. A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. BMC Genomics 7:206.