

RESEARCH ARTICLE

Open Access

# Methods of tagSNP selection and other variables affecting imputation accuracy in swine

Yvonne M Badke<sup>1</sup>, Ronald O Bates<sup>1</sup>, Catherine W Ernst<sup>1</sup>, Clint Schwab<sup>3</sup>, Justin Fix<sup>4</sup>, Curtis P Van Tassel<sup>5</sup> and Juan P Steibel<sup>1,2\*</sup>

## Abstract

**Background:** Genotype imputation is a cost efficient alternative to use of high density genotypes for implementing genomic selection. The objective of this study was to investigate variables affecting imputation accuracy from low density tagSNP (average distance between tagSNP from 100kb to 1Mb) sets in swine, selected using LD information, physical location, or accuracy for genotype imputation. We compared results of imputation accuracy based on several sets of low density tagSNP of varying densities and selected using three different methods. In addition, we assessed the effect of varying size and composition of the reference panel of haplotypes used for imputation.

**Results:** TagSNP density of at least 1 tagSNP per 340kb (~ 7000 tagSNP) selected using pairwise LD information was necessary to achieve average imputation accuracy higher than 0.95. A commercial low density (9K) tagSNP set for swine was developed concurrent to this study and an average accuracy of imputation of 0.951 based on these tagSNP was estimated. Construction of a haplotype reference panel was most efficient when these haplotypes were obtained from randomly sampled individuals. Increasing the size of the original reference haplotype panel (128 haplotypes sampled from 32 sire/dam/offspring trios phased in a previous study) led to an overall increase in imputation accuracy ( $I_A = 0.97$  with 512 haplotypes), but was especially useful in increasing imputation accuracy of SNP with MAF below 0.1 and for SNP located in the chromosomal extremes (within 5% of chromosome end).

**Conclusion:** The new commercially available 9K tagSNP set can be used to obtain imputed genotypes with high accuracy, even when imputation is based on a comparably small panel of reference haplotypes (128 haplotypes). Average imputation accuracy can be further increased by adding haplotypes to the reference panel. In addition, our results show that randomly sampling individuals to genotype for the construction of a reference haplotype panel is more cost efficient than specifically sampling older animals or trios with no observed loss in imputation accuracy. We expect that the use of imputed genotypes in swine breeding will yield highly accurate predictions of GEBV, based on the observed accuracy and reported results in dairy cattle, where genomic evaluation of some individuals is based on genotypes imputed with the same accuracy as our Yorkshire population.

**Keywords:** Genotype imputation, Pigs, Reference panel size

\*Correspondence: steibelj@msu.edu

<sup>1</sup>Department of Animal Science, Michigan State University, East Lansing, MI, USA

<sup>2</sup>Department of Fisheries & Wildlife, Michigan State University, East Lansing, MI, USA

Full list of author information is available at the end of the article

## Background

Recent advances in genotyping technology have facilitated the availability of high density genotyping platforms in many livestock species. High density platforms including several thousand single nucleotide polymorphisms (SNP) are available for cattle [1-3], chicken [4], sheep [5], and pig [6].

These platforms can be used to increase the efficiency and accuracy of breeding programs by implementing genomic selection [7,8]. Using SNP data to inform breeding decisions allows animal breeders to select breeding stock prior to the animals having progeny of their own, thereby accelerating genetic progress through shortened generation intervals [7,8].

Currently, genomic selection has been successfully implemented in dairy cattle based on genotypes from the Illumina BovineSNP50 chip [7]. In an effort to increase cost efficiency, the use of low density (tagSNP) genotyping platforms was exploited for dairy cattle [9,10]. If high density genotypes are imputed from tagSNP with high accuracy, the loss of reliability of predicted genomic breeding values is minimal [9-11]. High accuracy of imputed genotypes depends on the selection of tagSNP, as well as the composition and size of the reference panel of haplotypes used for imputation.

If close relatives of all imputation candidates are genotyped at high density, untyped markers can be recovered through linkage and segregation analysis [12], where haplotypes can be traced through generations of directly related individuals using the rules of Mendelian inheritance. However, in some species it may not be feasible to genotype a large proportion of the pedigree at high density. In that case a small panel of reference haplotypes can be used to impute all untyped markers by exploiting population-wide linkage disequilibrium (LD) [13,14]. This approach was initially proposed in human genome-wide association studies (GWAS) and has recently found application in plant [15] and animal breeding [10,11,16]. A combination of imputation based on segregation analysis and population-wide LD is currently being used in dairy breeding [9]. While combining both approaches will increase accuracy of imputation, eventually becoming the default method, cost-effective implementation of genomic selection in novel populations is likely to initially rely more on LD based imputation. Consequently, in this paper we will concentrate on LD based imputation by investigating tagSNP selection and haplotype reference panel construction.

Human geneticists have proposed a variety of approaches to select an optimal low density set of tagSNP to achieve cost efficient imputation in GWAS [17]. These approaches include statistical criteria based on a pairwise threshold of LD between SNP (i.e. [18]) and predictive ability, selecting tagSNP that provide the most accurate

prediction of all non-typed markers [19]. On the other hand, tagSNP sets used in livestock are mainly selected for equidistant spacing based on physical position along the genome, and high minor allele frequency (MAF) to ensure segregation (e.g. [1]).

Crucial to successful implementation of genotype imputation using population wide LD is the availability of a representative panel of reference haplotypes [20,21]. These panels are commonly built by genotyping a small number of trios or a larger number of relatively unrelated individuals. The overall goal in either case is to collect genotypes that can be accurately phased [22] into haplotypes representative of population frequencies. As a result, we began our study by genotyping and phasing a small number of trios in four US pig breeds [23] ( $N_{\text{Trios}} \sim 30$ ) and further enriching this panel for the Yorkshire breed with a set of high density genotypes from largely unrelated individuals ( $N_{\text{samples}} = 889$ ).

The objective of this study was to develop guidelines for the implementation of genotype imputation in livestock populations having little or no prior use of genome-wide marker-assisted-selection. First, we compared imputation accuracy resulting from three methods of tagSNP selection using Yorkshire pigs genotyped with a high density SNP set (Illumina PorcineSNP60). This includes a report on imputation accuracy of the recently developed commercially available 9K tagSNP set referred to as the GeneSeek Genomic Profiler for Porcine LD (GGP-Porcine, GeneSeek a Neogen Company, Lincoln, NE). Second, we assess accuracy of imputation based on an increasing number of reference haplotypes to inform the selection of an optimal reference panel of haplotypes. Finally, we discuss imputation accuracy as a function of chromosomal location and MAF of non-observed SNP.

## Methods

### Genotypes

High density genotypes for approximately 30 sire/dam/offspring trios were obtained and phased for each of four breeds of pigs (Duroc, Hampshire, Landrace, Yorkshire) in a previous study [23]. To ensure accurate phasing, the reference panel for imputation used in this study was the 128 haplotypes from the Yorkshire sire/dam pairs previously genotyped as parents in those trios. Animal protocols were approved by the Michigan State University All University Committee on Animal Use and Care (AUF# 03/09-046-00). The haplotypes of these animals are freely available at [https://www.msu.edu/~steibelj/JP\\_files/LD\\_estimate.html](https://www.msu.edu/~steibelj/JP_files/LD_estimate.html).

Detailed information about data cleaning procedures, descriptive statistics of LD, and correlation of phase between Yorkshire and other US pig breeds can be found in Badke et al. [23]. In addition, DNA samples were

collected from 920 Yorkshire pigs and sent to a commercial laboratory (GeneSeek, a Neogen Company, Lincoln, NE) to be genotyped on the Illumina PorcineSNP60 (Number of markers  $M=62,163$ ) Genotyping BeadChip (Illumina Inc.) [6]. Only animals with more than 90% genotype call rate were considered for analysis, resulting in 889 animals used as the testing panel for this study. All SNP included in the 128 haplotype Yorkshire reference panel were used for analysis. All data from this study is available at [https://www.msu.edu/~steibel/JP\\_files/imputation.html](https://www.msu.edu/~steibel/JP_files/imputation.html).

In our previous study [23] we reported breed specific LD and persistence of phase among breeds for Duroc, Hampshire, Landrace, and Yorkshire pigs. We found that persistence of phase between Yorkshire and the other breeds ranged between 0.42 and 0.57 for SNP spaced approximately 1MB apart [23]. As a result the amount of LD within the Yorkshire breed that could be recovered through haplotypes from another breed ranges between 0.18 and 0.33, such that adding haplotypes of a second breed to impute Yorkshire genotypes did not appear to be beneficial. For genomic selection, a simulation study conducted by de Roos et al. [24] found that persistence of phase between breeds needs to be much larger than the reported value between Yorkshire and any of the other three breeds to implicate any advantage for the use of mixed breed training panels. For this reason we decided to use only Yorkshire haplotypes in the reference panel for imputation in this paper.

### Genotype imputation and estimation of imputation accuracy

All imputations in this study were done using BEAGLE version 3.3.1 [13], a genotype imputation software that uses a reference panel of haplotypes to estimate phase and impute missing genotypes in a set of unrelated individuals. Beagle was run separately for each chromosome using 128 reference haplotypes from the trio design [23] (phased file) to phase and impute genotypes of the 889 un-phased testing animals. All SNP, except tagSNP, were masked as missing in the testing set. Beagle was run for ten iterations of the phasing algorithm, drawing four samples per iteration. Previous results from another study [16], as well as a short experiment conducted in this study (data not shown) found no increase in imputation accuracy when the number of iterations or samples per iteration were increased. The output files from BEAGLE contained the most likely imputed genotypes (AA, AB, BB), posterior genotype probabilities ( $P(AA)$ ,  $P(AB)$ ,  $P(BB)$ ), and posterior expected allelic dosage of the B allele derived from the posterior genotype probabilities (i.e.  $0 * P(AA) + 1 * P(AB) + 2 * P(BB)$ ) [25].

Imputation accuracy was estimated using three different measures that reflect different influences of MAF

and error counting. The proportion of correctly imputed alleles was computed as

$$IA = 1 - \frac{\sum_{i=1}^M \sum_{j=1}^{N_i} |g_{ij} - \hat{g}_{ij}|}{2 * \sum_{i=1}^M N_i} \quad (1)$$

where  $g_{ij}$  is the observed allelic dosage of SNP  $i$  in individual  $j$ ,  $\hat{g}_{ij}$  is the corresponding posterior expected allelic dosage obtained from BEAGLE output,  $M$  is the total number of imputed SNP, and  $N_i$  is the number of individuals with called genotypes for SNP  $i$ . This overall measure of imputation accuracy can be further decomposed into

SNP-specific accuracy ( $IA_i = 1 - \frac{\sum_{j=1}^{N_i} |g_{ij} - \hat{g}_{ij}|}{2 * N_i}$ ) and animal specific accuracy ( $IA_j = 1 - \frac{\sum_{i=1}^{M_j} |g_{ij} - \hat{g}_{ij}|}{2 * M_j}$ ). This measure of imputation accuracy will be biased upwards, especially for SNP with low MAF, because even if imputation ignores LD information and is based solely on allele frequency, the major allele would be correctly imputed for a large proportion of genotypes [15,16]. As tagSNP density decreases imputation accuracy of rare alleles further decreases as rare haplotypes become harder to identify due to longer sequences of SNP missing [15]. Estimating the total percentage of correctly imputed alleles for SNP with low MAF will be biased due to the large number of correctly imputed major alleles masking the small number of misspecified minor alleles, which can be overcome through the use of a more sensitive measure of accuracy for these SNP [15]. In addition, if individuals carrying the minor allele are not correctly identified and their phenotype cannot be matched for GWAS this relatively small proportion of incorrectly imputed alleles will further decrease power. A variety of measures have been introduced to obtain estimates of imputation accuracy unbiased by MAF [15,16,26]. We estimated the proportion of correctly imputed alleles adjusted for MAF using the formula presented by Hayes et al. [16]:

$$IA_{MAF} = \frac{IA - IA_{Freq}}{1 - IA_{Freq}} \quad (2)$$

where  $IA$  is computed as described in equation (1) and  $IA_{Freq}$  is the accuracy of imputation based on genotypic frequencies estimated as:

$$IA_{Freq} = p(AA)_{ref} * p(AA)_{val} + p(AB)_{ref} * p(AB)_{val} + p(BB)_{ref} * p(BB)_{val} \quad (3)$$

where  $p(AA)_{ref}$ ,  $p(AB)_{ref}$ , and  $p(BB)_{ref}$  are the observed frequencies for genotypes AA, AB, and BB for SNP  $i$  in the reference haplotypes and  $p(AA)_{val}$ ,  $p(AB)_{val}$ , and

$p(BB)_{val_i}$  are the predicted genotypic frequencies in the testing population for SNP  $i$ .  $IA_{Freq}$  can be interpreted as the expected probability of correctly imputing a genotype in the testing population by assigning a randomly sampled genotype from the haplotypes in the reference panel. This measure was computed on a SNP-wise basis and averaged across all SNP. To account for a slightly different number of genotypes observed within each SNP (due to missing at random) the average was obtained by weighting the accuracy of each SNP by the number of individuals with observed genotypes within each SNP.

Alternatively, another measure of imputation accuracy robust to MAF is the squared correlation between the observed and imputed allelic dosage [15]. The correlation was obtained on a SNP by SNP basis using the correlation function in R [27]. SNP wise correlation measures were weighted by the number of available observations within the SNP to obtain an overall average imputation accuracy.

#### Methods of tagSNP selection

TagSNP were selected using three approaches: 1) evenly spaced based on physical position, 2) based on minimum pairwise LD with non-tagSNP (statistical selection), and 3) based on marker predictive ability to accurately impute non-observed SNP genotypes (predictive selection).

To select evenly spaced SNP the total length of each chromosome was partitioned into segments corresponding to the total number of tagSNP to be selected. Then, within each segment the SNP closest to the segment center was identified and added as a tagSNP. If a given segment was empty, no tagSNP was selected in that segment.

To implement a statistical search for tagSNP [19] we used the freely available software package FESTA [18]. FESTA performed a greedy search, where each SNP  $i$  was either an element of the tagSNP set or in LD higher than a threshold ( $r_t^2$ ) with an existing element of the tagSNP set. FESTA was run repeatedly for increasing  $r_t^2$  ranging from 0.1 to 0.9 in 0.1 increments using estimates of LD based on 128 reference haplotypes.

To implement predictive tagSNP selection, we applied the following forward search algorithm: First, we split the 64 Yorkshire reference animals into a randomly sampled set of 10 individuals (training set) and 54 individuals (reference haplotypes). Second, all SNP except one tagSNP in the training set were masked and imputed using the reference haplotypes. Third, accuracy of all imputed SNP was estimated and saved. Steps two and three were then repeated until all estimates of imputation accuracy were available for all potential tagSNP (at first, the potential tagSNP are all SNP on the chromosome). Fourth, the SNP that yielded the highest average accuracy of imputation among those not already chosen as tagSNP was selected as a new tagSNP. Steps two through four were repeated until the maximum number of tagSNP or a target imputation

accuracy were reached. Because of the high computational demand of this methodology, this approach was only applied to the smallest available chromosome (SSC18), selecting tagSNP from 786 candidate SNP.

Concurrent to this research, a set of 9390 tagSNP (Release Date: April 2012) was assembled by GeneSeek (Lincoln, NE) for the development of a commercial platform for low density genotyping in swine. This assay has been marketed as the GeneSeek Genomic Profiler for Porcine LD (GGP-Porcine; GeneSeek, Lincoln, NE). After production, the GGP-Porcine contains approximately 8500 tagSNP (Jeremy Walker, personal communication). TagSNP covering the entire genome were selected based on MAF in 13 commercial lines of pigs represented by four breeding companies and four purebred populations. The MAF were provided by the breeding companies (identified simply as company A, B, C, and D). The number of lines provided by these companies were 1, 1, 4, and 7. Additional estimates of MAF used to identify tagSNP were obtained from our previous study [23] of four pure breeds: Duroc, Hampshire, Landrace, and Yorkshire. The freely available SNPspace software (C.P. Van Tassell, unpublished data) was used to select tagSNP. SNPspace was initially developed to select SNP for the Illumina BovineSNP50 beadchip [2]. The conceptual framework of SNPspace is briefly described in that study [2], but additional features have been added since that time. Relative weights on lines or breeds of pigs ranged from 0.00625 to 0.25. SNPspace is based on a greedy algorithm, where SNP scores account for breed or line specific MAF, region of the genome, and position of SNP relative to previously selected tagSNP. Density of tagSNP was doubled within 5 Mbp of the chromosomal extremes, which has been shown to improve average accuracy of imputation compared to tagSNP evenly spaced across the entire chromosome [1,28].

#### Increasing reference panel size

To assess the effect of the number of reference haplotypes on imputation accuracy, we split the available sample of 889 Yorkshire pigs into two groups: 1) a 200 animal testing panel, and 2) a 689 animal set of supplemental reference sires. Assignment to the two panels was random.

To obtain imputation accuracy for a decreased set of reference haplotypes, we split the original 128 reference haplotypes obtained from 64 Yorkshire animals into two groups of 64 reference haplotypes (corresponding to 32 animals) and estimated average imputation accuracy in the 200 animal testing set. Then, we split the two groups of 64 reference haplotypes further into two groups of 32 reference haplotypes and obtained four estimates of imputation accuracy that were averaged into a single measure.

Subsequently, we compared imputation accuracy using trio based reference panels to imputation accuracy based

on randomly sampled reference panels. To this end, we randomly sampled 16 animals from the 689 animal supplemental reference set and continued to add individuals at random to obtain reference sets of 24, 32, 48, 64, 96, 128, 256, and 512 animals. Each of these sets was phased individually using BEAGLE [13] and then those haplotypes were used as reference panel to impute the 200 testing animals.

Finally, we assembled reference panels of haplotypes combining the original 128 haplotypes from trios, with an increasing number of supplemental reference sires. To form these reference panels 64, 128, 192, and 448 supplemental reference sires were randomly selected and phased using the trio haplotypes as a reference panel. Both, the trio reference haplotypes and an increasing number of supplementary reference haplotypes were then used to impute the 200 animal testing set.

Because imputation accuracy was constant across chromosomes (see Results, section 3.1) we conducted this experiment on chromosome SSC14, a medium sized chromosome that has uniform coverage of SNP across its length. We expect results to extrapolate to all other chromosomes.

## Results

### Comparison of methods for tagSNP selection

Due to the high computational demand, we initially performed a comparison of methods for tagSNP selection only on the smallest chromosome (SSC18). Statistical tagSNP selection requires fixing an  $r^2$  threshold ( $r_t^2$ ). Setting  $r_t^2 = 0.2$  resulted in the selection of 165 tagSNP, which produced imputation accuracy of 0.936. Increasing  $r_t^2$  to 0.3, led to a panel of 235 tagSNP and an increased imputation accuracy of 0.956. In comparison, imputation accuracy based on 165 and 235 tagSNP selected for predictive ability was 0.93 and 0.945, respectively. Direct comparison to tagSNP sets selected for even spacing is more difficult because of empty intervals, for which no tagSNP were selected, resulting in smaller than targeted tagSNP sets. The evenly spaced tagSNP sets closest in size to 165 and 235 tagSNP were as expected slightly smaller (161 and 224 tagSNP), and the resulting imputation accuracies were slightly lower than those obtained using the other sets (0.92 and 0.941, respectively). As expected, imputation accuracy increased with increasing densities of tagSNP regardless of the selection method (Figure 1). Statistically selected tagSNP performed slightly better than both, predictive and evenly spaced tagSNP (Figure 1), but all three methods resulted in similar imputation accuracy. Selection of tagSNP using predictive ability required an at least 500-fold increase in computation time for SSC18 compared to statistical and evenly spaced selection. However, results of imputation accuracy indicate that predictive tagSNP did not yield significantly

higher imputation accuracy compared to tagSNP selected by other methods. Therefore, only statistical and evenly spaced tagSNP were selected in an exhaustive evaluation of imputation accuracy across all autosomes (Figure 2).

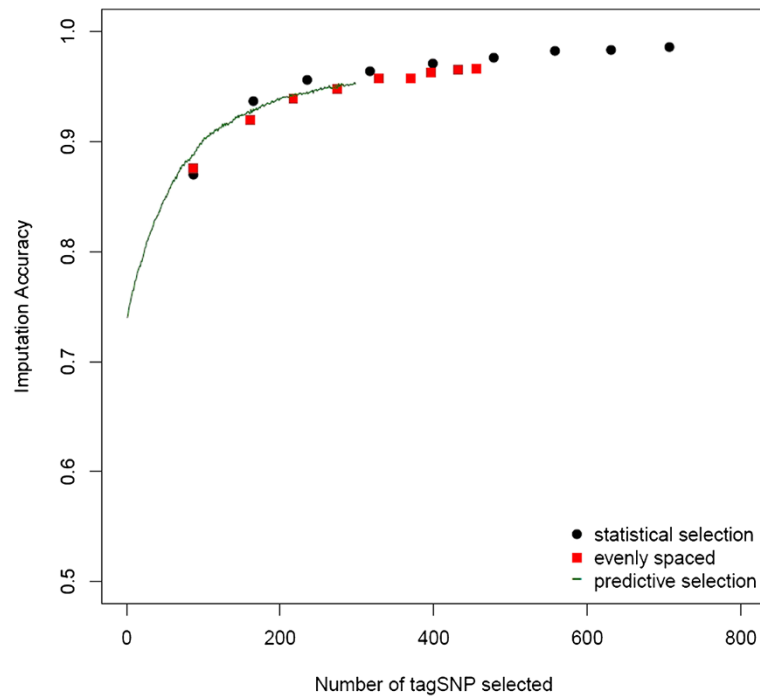
When imputing across all autosomes, as observed on SSC18, imputation accuracy using statistically selected tagSNP was slightly higher than that using evenly spaced tagSNP (Figure 2). In particular, to attain imputation accuracy of 0.95, 7036 statistically selected tagSNP were necessary ( $r_t^2 = 0.3$ ). In comparison, 10540 evenly spaced tagSNP were necessary to reach similar imputation accuracy. Imputation accuracy was virtually uniform across chromosomes ranging from 0.92 to 0.94 for  $r_t^2 = 0.2$  and from 0.94 to 0.96 for  $r_t^2 = 0.3$ .

We computed imputation accuracy based on 7323 tagSNP from the original list of 9K tagSNP provided by GeneSeek that passed quality control in this study ( $MAF > 0.05$ ,  $CallRate > 0.9$ , assembled to an autosome under map build10) resulting in imputation accuracy of 0.951 with a SNP-wise 95% highest posterior density interval equal to [0.84, 1] (Figure 2). Accuracy of imputation using the commercial tagSNP was similar to that obtained using statistically selected SNP, at comparable density ( $r_t^2 = 0.3$ ,  $M_{tagSNP} = 7036$ ). The advantage of the proposed commercial platform is that it is not based on population specific LD, thereby making it applicable across swine populations. For this reason all subsequent results of imputation accuracy will be based on the tagSNP element of the Genomic Profiler for Porcine LD.

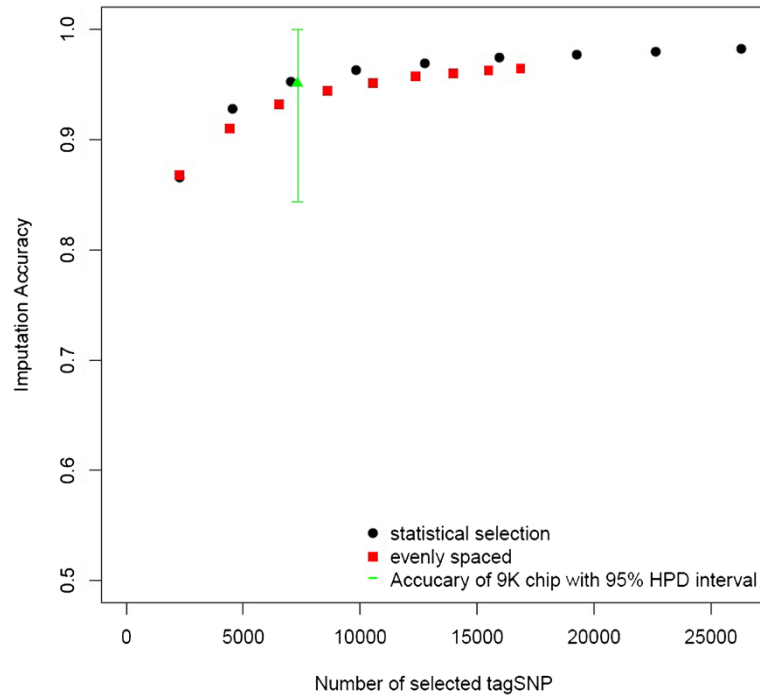
### Imputation accuracy using the commercial 9K tagSNP set

To assess accuracy of imputation as a function of chromosomal location we plotted imputation accuracy of each individual SNP versus chromosomal position (Figure 3). SNP within 5% of the chromosomal extremes had on average slightly lower imputation accuracy (0.949) than the 10% in the center of the chromosome (0.972). As mentioned before, this property of imputation accuracy has previously been observed for other low density sets [1,28] and was anticipated during the tagSNP set design. Based on these reports [1,28], the density of tagSNP was approximately doubled within 5 Mbp of the chromosomal ends in the commercial 9K tagSNP set.

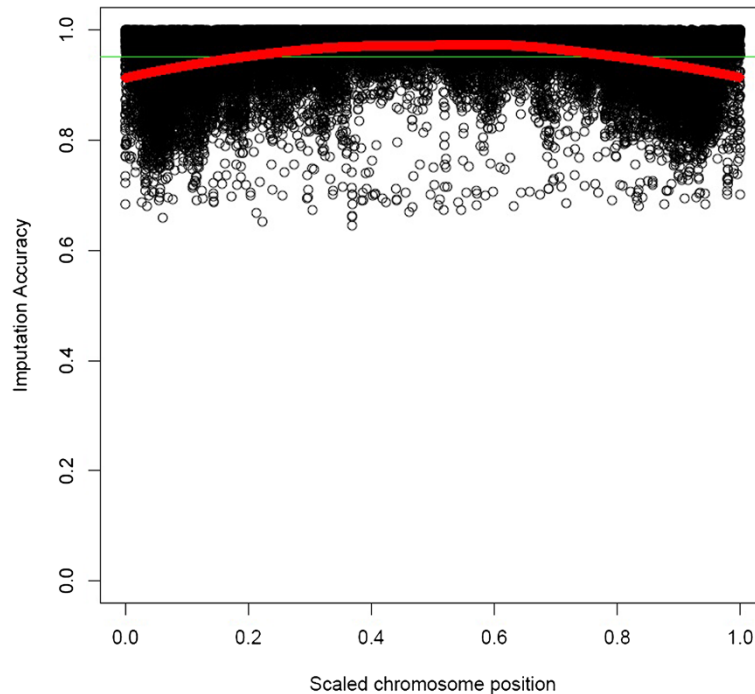
Animal-wise imputation accuracy ( $IA_i$ ) averaged 0.951 but the corresponding highest posterior density interval ([0.917, 0.978]) was shorter than that observed for SNP-wise accuracy. Overall, all but 12 animals had imputation accuracy  $> 0.90$  and 551 animals (62%) had imputation accuracy above 0.95. Also, seven of the animals had a dam, sire, or grand-sire in the reference panel [23], which resulted in on average higher accuracy of imputation in these animals (0.959). A group of 15 animals was identified with consistently low imputation accuracy (i.e.  $< 0.91$  for 9K tagSNP) across all sets of tagSNP selected. An



**Figure 1 Imputation accuracy based on tagSNP selected using 3 different methods.** Average imputation accuracy (IA) as a function of the number of tagSNP selected using three methods of tagSNP selection for SSC18: 1) evenly spaced (red square), 2) statistical selection (black circle), or 3) predictive selection (green line).



**Figure 2 Imputation accuracy using evenly spaced or statistically selected tagSNP.** Average imputation accuracy (IA) as a function of the number of tagSNP selected for: 1) even spacing (red square), or 2) statistical selection (black circle) across all autosomes. Imputation accuracy for 7323 tagSNP from the commercial 9K tagSNP set (green triangle) with 95% highest posterior density interval.

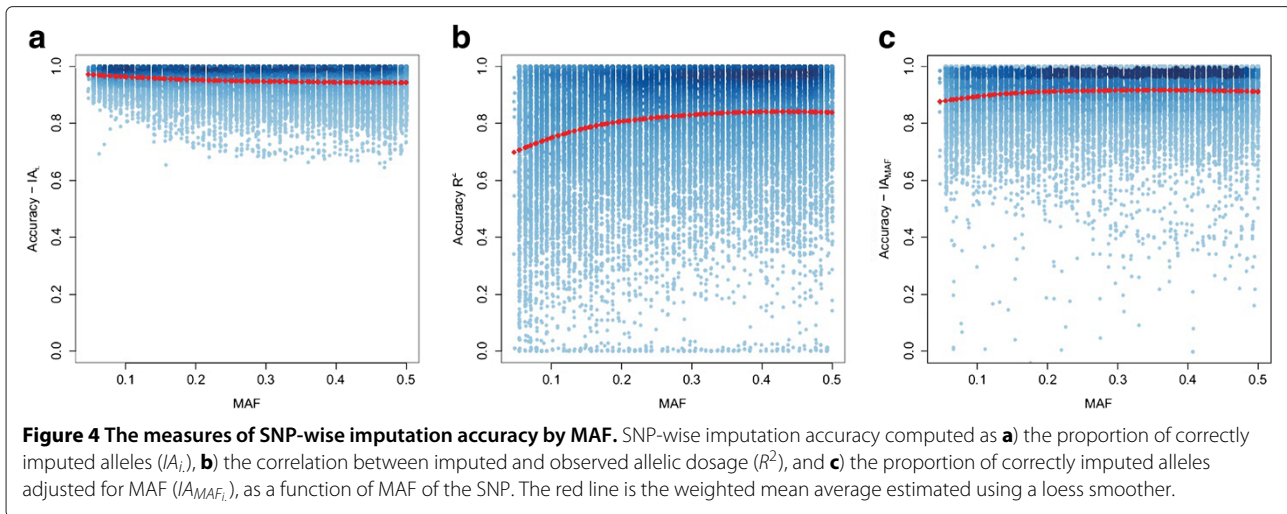


**Figure 3 SNP-wise imputation accuracy by chromosomal location.** SNP-wise imputation accuracy ( $IA_i$ ) vs. the scaled chromosomal location of the SNP. The red line is the weighted mean average estimated using a loess smoother [29], and the green line represents average imputation accuracy ( $IA = 0.951$ ).

ongoing research project in our laboratory investigating breed composition, identified all of the 15 low accuracy individuals as potentially having mixed breed ancestry (Yijian Huang, unpublished data). Further assessing the pedigree of these 15 animals, we found that nine of them were imported to the US, which could result in a slightly different haplotype composition and the observed low accuracy of imputation, when only American Yorkshire pigs had been used as reference. Another three animals of the remaining six US Yorkshires with low imputation accuracy were identified as a family (sire, two offspring), such that the observed low accuracy in the offspring is likely a result of the mixed breed ancestry of their sire.

As noted before, to assess the effect of MAF of imputed SNP on imputation accuracy required adjusting estimates of imputation accuracy for MAF. Imputation accuracy as a function of MAF is presented in Figure 4, where imputation accuracy was estimated as a) proportion of alleles correctly imputed, b) coefficient of determination ( $R^2$ ) between observed and imputed allelic dosage [15,26], and c) proportion of alleles correctly imputed adjusted for MAF [16]. The red line in all plots represents the weighted mean average estimated using a loess smoother [29]. Loess consists of fitting smooth piecewise polynomial regressions to local subsets of data and it is widely used in normalization of micro-array experiments [30]. At first

inspection, it can be seen that accuracy estimated as the proportion of correctly imputed alleles (Figure 4a) is highest for low frequency alleles and exhibits a small decrease as MAF increases. However, the observed high proportion of correctly imputed alleles in SNP with low MAF is based on the fact that high frequency alleles can be imputed with high accuracy even if imputation is solely based on allele frequency [15,16]. For this reason, we computed  $R^2$  and the proportion of correctly imputed alleles adjusted for MAF that provide estimates of imputation accuracy unbiased by allele frequency. In other words, these measures are indicative of the performance of the imputation algorithm in comparison to a baseline imputation based on genotypic frequencies [15,16]. When imputation accuracy is adjusted for MAF, estimated accuracy is generally higher for intermediate allele frequencies ( $MAF \sim 0.5$ ) and declines as MAF decreases (Figure 4 b/c). Average imputation accuracy considering only the added benefit of the imputation algorithm was lower ( $IA_{MAF} = 0.91$ ,  $R^2 = 0.81$ ) than the total proportion of correctly imputed alleles ( $IA = 0.951$ ). The difference between the proportion of correctly imputed alleles adjusted for MAF (Figure 4c) and estimates of  $R^2$  (Figure 4b) can be explained by the difference in error counting between these measures. While the proportion of correctly imputed alleles adjusted for MAF is obtained by counting the total number of wrongly



imputed alleles,  $R^2$  is obtained from the squared difference in imputed and observed alleles, thereby more heavily penalizing large differences between observed and imputed allelic dosage.

#### Effect of numbers of reference haplotypes on imputation accuracy

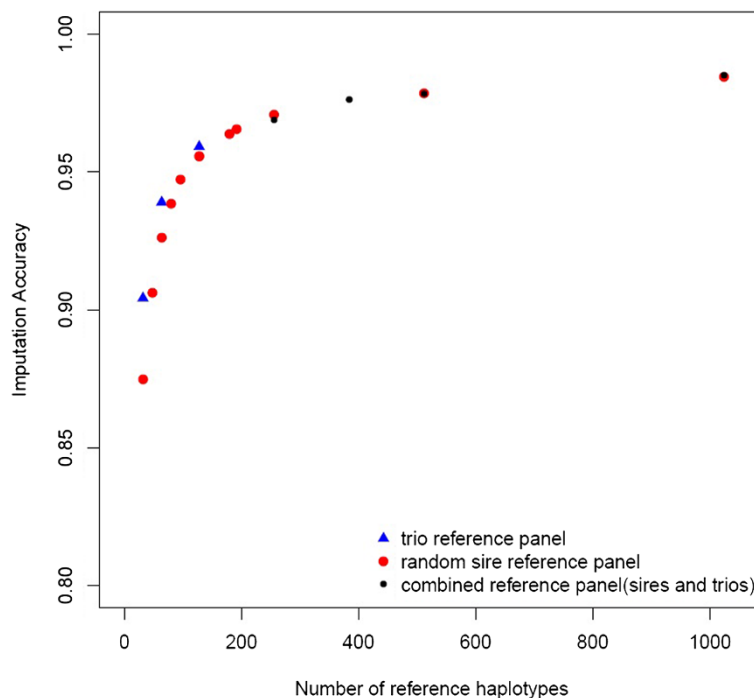
For all previous analyses in this paper we imputed genotypes of 889 individuals across all autosomes using a reference panel of 128 Yorkshire haplotypes obtained from a sire/dam/offspring genotyping design [23], phased with higher accuracy [22]. Reducing the number of imputation animals from 889 to 200 had no impact on the observed imputation accuracy. Imputation accuracy using all 128 haplotypes from the original reference panel was 0.959 on SSC14, which reduced to 0.939 when 64 haplotypes were used, and further to 0.904 when imputation was based on 32 haplotypes (Figure 5). Therefore, imputation accuracy larger than 0.90 can be obtained using the commercial 9K tagSNP set with a reference panel of only 32 haplotypes, given that these haplotypes were phased at high accuracy.

We further investigated if it is necessary to obtain reference haplotypes from a trio design, or if accuracy can be replicated using a reference panel of randomly sampled individuals genotyped at high density. In comparison to imputation accuracy obtained using a trio reference panel, imputation accuracy based on 32 and 64 reference haplotypes derived from selected sires was slightly lower (0.875 and 0.926, respectively). However, accuracy from 128 reference haplotypes obtained from 64 randomly sampled individuals was 0.955, which is practically identical to results obtained using 128 reference haplotypes from trios. Therefore, if the reference panel of haplotypes is composed of more than 128 haplotypes, there is no longer an advantage in using haplotypes obtained from a trio design. Alternatively, the cost of assembling panels of 32,

64, and 128 reference haplotypes obtained from a trio design involves the same genotyping cost as assembling panels of 48, 96, and 192 haplotypes obtained from randomly sampled individuals. This is due to the fact that in a trio design the offspring haplotypes are not used as part of the reference panel, since they are identical to the parents transmitted haplotypes. Imputation accuracies for 48, 96, and 192 reference haplotype panels from randomly sampled individuals were estimated to be 0.906, 0.947, and 0.965, respectively, which is either equivalent or higher than accuracy of imputation obtained using the cost equivalent trio based reference panels (Figure 5). In addition, we compared imputation accuracy from reference haplotypes of either 64 randomly selected individuals or the 64 oldest individuals and found no difference in imputation accuracy (0.956, 0.953 respectively). Consequently, if no reference panel of haplotypes is available for a population, according to the results of this study, it would be most cost efficient to assemble high density haplotypes of randomly sampled individuals.

Previous research has indicated an increase in imputation accuracy can be expected as the number of available reference haplotypes increases to a certain point [20,21]. We added randomly selected individuals to the reference panel and obtained 256, 512, and 1024 reference haplotypes. These panels resulted in average accuracy of imputation of 0.971, 0.978, and 0.985 respectively (Figure 5). Imputation accuracy only marginally increased when more than 256 haplotypes were used as reference panel for imputation (up to 1.4% gain). Additionally, we assessed accuracy of imputation from reference panels composed of the original 128 reference haplotypes from a trio design and an increasing number of randomly sampled individuals added to that panel. In this case, imputation accuracy based on reference panels with 256 and 512 haplotypes was 0.969 and 0.978 respectively, which is





**Figure 5 Effect of number of reference haplotypes on imputation accuracy.** Average imputation accuracy (IA) as a function of the number of haplotypes in the reference panel used for imputation. Imputation accuracy was estimated for reference panels composed of haplotypes from a trio design (blue triangle), reference panels composed of haplotypes from randomly sampled sires (red circle), and reference panels composed of both haplotypes from a trio design and haplotypes from randomly sampled sires (black circle).

virtually identical to results obtained using reference panels solely from randomly sampled individuals (Figure 5).

In addition to assessing the effect of an increased number of reference haplotypes on average accuracy, we also investigated how it affects individual SNP with different MAF and physical location. We found that as the size of the reference panel increases, imputation accuracy (quantified as  $R^2$ ) improved more markedly for SNP with MAF below 0.1, such that when the size of the reference panel is increased from 256 haplotypes to 512 haplotypes the increase in accuracy for SNP with MAF below 0.1 was on average 0.06 points, while for all other SNP the increase was only 0.02 points (Additional file 1: Figure S1). When imputation was based on 1024 reference haplotypes imputation accuracy appears to be uniform across allele frequencies. Similarly, we observed that imputation accuracy (proportion of correctly imputed alleles) for SNP located in the 10% chromosomal extremes (5% on either side) could be improved through an increase in the number of reference haplotypes (Additional file 2: Figure S2). A reference panel containing 512 haplotypes was necessary to obtain maximal imputation accuracy ( $IA = 0.99$ ) for SNP located in the chromosomal center, while SNP in the chromosomal extremes were imputed with accuracy of only 0.97, even when the number of

reference haplotypes was doubled (1024 reference haplotypes). Imputation accuracy observed in SNP located in the chromosomal extremes was more than 0.02 accuracy units lower than the average imputation accuracy of all remaining SNP irrespective of the reference panel size.

## Discussion

### Methods for tagSNP selection

Current algorithms for genotype imputation exploit population-wise LD [13,14], familial LD from identity by descent [31], or a combination of both [32] to infer unobserved genotypes conditional on tagSNP information. Virtually all methods for tagSNP selection aim at identifying tagSNP that carry the maximum amount of information to impute unobserved markers. This is attained by either directly quantifying the tagSNP ability to predict non-typed SNP (predictive tagSNP selection) or indirectly by selecting tagSNP in high pairwise LD with non-tagSNP (statistical tagSNP selection) [17].

A goal of this study was to select a minimal set of tagSNP that would yield acceptable accuracy of imputation of non-tagSNP [10,28]. Since genotype imputation utilizes information about the structure of LD to infer non-observed SNP, we expected that tagSNP sets selected based on LD information, such as statistical and predictive

tagSNP selection, would yield higher accuracy of imputation than tagSNP selected based solely on their physical location. In addition, we expected that directly assessing the ability of each tagSNP to predict non-observed SNP (predictive selection) would yield an improvement in imputation accuracy compared to tagSNP selected purely based on pairwise thresholds of LD (statistical selection).

We found that at the lowest examined tagSNP density (1 tagSNP per Mb) accuracy of imputation was below 0.87 irrespective of the method of tagSNP selection and that at least 2 tagSNP per Mb were necessary to increase accuracy to at least 0.91. Accuracy of imputation increased as tagSNP density increased, reaching a plateau accuracy of approximately 0.98 when tagSNP were spaced at an average distance of less than 125kb with negligible increases beyond such density. Our results compare well to those of Weigel et al. [33], where randomly selected tagSNP at an approximate density of 300kb were necessary to obtain accuracy larger than 0.90 in the US Jersey cattle population using a similar type of imputation. In our study, imputation accuracy of approximately 0.95 was obtained using between 7000 (average tagSNP spacing of 340kb) and 10000 tagSNP (average tagSNP spacing of 230kb), depending on the method of tagSNP selection.

As expected, predictively and statistically selected tagSNP did yield higher accuracy of imputation than evenly spaced tagSNP, but we found no difference in imputation accuracy between tagSNP sets selected statistically or based on predictive ability. Comparing 300 tagSNP selected using predictive ability to 317 tagSNP obtained using statistical selection ( $r_t^2 = 0.4$ ) on SSC18, we observed the same imputation accuracy ( $IA = 0.95$ ). However, the two sets are qualitatively different. For instance, the 300 predictive tagSNP only provide statistical coverage ( $r^2 \leq 0.4$ ) to 37% of non-tagSNP. The tagSNP sets also have on average different MAF ( $MAF_{predictive} = 0.30$ ,  $MAF_{statistical} = 0.27$ ). We attribute the equivalence in imputation accuracy of two different tagSNP sets to the extent of LD observed across the genome in Yorkshire pigs ( $r^2 = 0.16$  at 1 Mb) [23]. Under these conditions, precision of estimates of individual tagSNP imputation accuracy is likely compromised by collinearity, making selection of a single best predictive tagSNP at each step of the forward search complicated [34]. For example, the initial step of the forward search for predictive tagSNP resulted in six SNP with predictive ability within 0.002 accuracy units of each other. Each of these SNP could have been selected as a starting point of the greedy search, resulting in different sets of tagSNP selected. Furthermore, the implemented predictive forward search requires  $\frac{M_i(M_i+1) - M_{ti}(M_{ti}+1)}{2}$  imputation operations per iteration step compared to only two with statistical selection, where  $M_i$  is the number of SNP per chromosome and  $M_{ti}$  is the number of selected tagSNP on that chromosome.

Consequently, even though both methods result in different tagSNP sets, statistical selection is a computationally efficient proxy for predictive tagSNP selection when moderate LD between consecutive markers is present.

We show that tagSNP sets strictly selected for even spacing are slightly outperformed by statistical or predictive tagSNP selection. However, it is possible to enhance the performance of evenly spaced tagSNP through a few simple measures. TagSNP with high MAF seem to be advantageous for genotype imputation (predictive tagSNP selection seemed to favor tagSNP with high MAF) and their likelihood to segregate across populations will ensure that they carry information for imputation in various populations. This has been exploited previously in cattle for the assembly of the 3K platform [28], as well as in newer tagSNP sets aimed to further increase imputation accuracy [1]. In addition to selecting evenly spaced tagSNP with high MAF, an increase in accuracy can be obtained by increasing tagSNP density in the chromosomal extremes [1]. The success of these enhancements of evenly spaced tagSNP is evident in the imputation accuracy we report using the commercial 9K set ( $M_{tagSNP} = 7323$ ,  $IA = 0.951$ ), which is similar to results we found for statistical tagSNP sets for thresholds  $r_t^2 = 0.3$  ( $M_{tagSNP} = 7036$ ,  $IA = 0.952$ ). In addition, although the recently released commercially available chip has approximately 10% fewer tagSNP than the original 9K tagSNP list that was used for this analysis, our conclusions regarding imputation accuracy are likely to uphold, due to the fact that we based our analysis on a set of only 7323 tagSNP, which should be representative of the number of commercial tagSNP that will pass quality control in future study samples.

In summary, efficient tagSNP selection based on MAF and physical location is feasible and more flexible than statistical tagSNP selection. Selecting evenly spaced tagSNP with high MAF requires knowledge of the physical location of the SNP and the MAF across populations of interest, while statistical tagSNP selection requires knowledge of the LD structure, and would be population specific. As a result, selecting a tagSNP set with high MAF and an increased density in the chromosomal extremes is more versatile than tagSNP sets selected for predictive ability or based on statistical criteria while yielding the same accuracy of imputation. In addition, the tagSNP set selected based on physical location and MAF is expected to be useful for imputation as long as the 60K chip is being used for genomic selection, because we do not expect selection to alter LD or MAF of selected SNP in any particular way. If such tagSNP sets will be used across multiple closely related populations it will be necessary to include a number of SNP that will be specific to a subset of populations. In the case of the 9K tagSNP set more than 9000 tagSNP were selected based on MAF across several populations and physical location of the SNP, but only 7323 of these

SNP passed quality editing for the Yorkshire data in this study.

### Factors affecting imputation accuracy

Accuracy of imputation is affected by several factors including the selection and density of tagSNP as detailed above, the MAF and the physical location of the imputed SNP, as well as the size and composition of the reference panel.

When evaluating imputation accuracy as a function of the tagSNP selection method and density we have focused on average accuracy as a measure of overall performance. Assessing the average accuracy of imputation is a good indicator of the performance of imputed genotypes, when all genotypes are used simultaneously to obtain a global measure. Such a measure could be prediction of GEBV, which would be based on all SNP simultaneously, such that a small number of wrongly imputed SNP is unlikely to greatly affect the accuracy of prediction. Alternatively, some applications of imputed high density genotypes may require high accuracy across all SNP. One example would be GWAS based on imputed genotypes. For GWAS, SNP associations are assessed on a SNP by SNP basis, such that wrongly imputed alleles for low frequency SNP are more likely to cause bias in the estimated association, especially since phenotypes of interest are suspected to be associated with low frequency alleles [20].

One of the factors directly related to the individual SNP imputation accuracy, is the allele frequency of that particular SNP. To investigate imputation accuracy as a function of MAF we used two measures of imputation accuracy that were unbiased by MAF (i.e.  $IA_{MAF}$ ,  $R^2$ ). The adjusted proportion of correctly imputed alleles ( $IA_{MAF}$ ) and the correlation between observed and imputed allelic dosage ( $R^2$ ) are scaled differently, such that the observed accuracy differs as a function of scale, but the comparative difference in imputation accuracy as a function of MAF can be observed using either of the two accuracy measures. We found that estimates of imputation accuracy adjusted for MAF ( $R^2$ ,  $IA_{MAF}$ ) are lower ( $R^2 = 0.73$ ,  $IA_{MAF} = 0.89$ ) for SNP with MAF below 0.1, compared to SNP with MAF above 0.1 ( $R^2 = 0.82$ ,  $IA_{MAF} = 0.91$ ), which has been previously noted by Hayes et al. [16] reporting results of genotype imputation in sheep and Hickey et al. [15] in lines of maize.

Another factor relating to individual SNP imputation accuracy is the physical location of the SNP. Previous studies designing low density genotyping platforms have pointed out the need to increase coverage of tagSNP in the chromosomal extremes due to difficulties in correctly imputing SNP located in those regions [1,28]. In the commercial 9K tagSNP set the density of tagSNP within 5Mbp of the chromosomal extremes was approximately doubled to aid imputation accuracy. We found that

imputation accuracy using the 9K commercial tagSNP was still slightly lower in the extreme regions (0.949) when compared to the chromosome center (0.972). The effect however was alleviated in comparison to an equally spaced tagSNP set of comparable density, where the average imputation accuracy in the chromosomal extremes was only 0.89.

We found a group of 15 animals that produced consistently low imputation accuracy ( $IA \leq 0.90$ ), compared to all remaining animals ( $IA = 0.951$ ). Nine of these animals were identified as imports, such that the observed low accuracy of imputation is likely a result of differences in haplotype frequencies between the US Yorkshire population that was used as reference for imputation, and the population(s) from which these animals originated. The remaining six animals were all identified as having potentially mixed breed ancestry based on results of a concurrent research project in our laboratory (Yijian Huang, unpublished data). In addition, we can infer from these results that if a population contains heterogeneous sub-populations, such as a large number of imported animals or animals with cross-bred ancestry, imputation accuracy will be decreased if this sub-structure is not accounted for when sampling reference haplotypes.

We found that increasing the number of reference haplotypes led to an increase in average imputation accuracy. In addition to the number of reference haplotypes, their average relatedness to the imputation candidates [15,16,35,36], as well as accurate phasing of these haplotypes [37] directly affect the resulting accuracy of imputation. In this paper, we assessed the effect of phasing accuracy and the number of reference haplotypes. Previous research comparing phasing accuracy of unrelated or randomly sampled individuals and trio designs (sire/dam/offspring), found that genotypes from trios can be phased with higher accuracy [22]. The initial reference panel available in this study was composed of the haplotypes of sire/dam pairs from a previous sample of trios that were unrelated for at least two generations and therefore sampled to efficiently cover the Yorkshire population [23]. We found that for haplotype panels composed of 64 or less haplotypes imputation accuracy was higher when these haplotypes were obtained from the trio design rather than a random sample of individuals (Figure 5). This advantage of the trio design is likely due to the superior phasing accuracy as well as the sampling strategy used to obtain these samples. However, adjusting sample size for the increased genotyping cost in a trio design, we observed that imputation accuracy was equal or higher when imputation was based on haplotypes obtained from randomly sampled individuals instead of trio reference haplotypes (Figure 5). Therefore, we conclude that if no reference panel is available in a population the most cost efficient method for reference panel

construction is genotyping a random sample of individuals across the population.

Next we assessed imputation accuracy as a function of increasing reference panel size. We found that a reference panel of 256 to 512 reference haplotypes is sufficient to obtain imputation accuracy of  $IA = 0.97$ . If the size of the reference panel is increased beyond 1024 haplotypes ( $IA = 0.985$ ) any further gain in imputation accuracy appears to be very small. A similar type of response has been observed in human genotype imputation [21]. This relatively small number of reference haplotypes necessary to obtain high imputation accuracy ( $IA = 0.97$ ) is likely due to the relatively small effective population size of the Yorkshire population ( $N_e = 113$ , [38]), and consequently high average LD even at decreased tagSNP density.

After determining that increasing the size of the reference panel would increase accuracy of imputation, we assessed whether accuracy would differ as a function of reference panel composition. In general, when imputation experiments are conducted the older animals are used as reference panel, while the younger animals serve as imputation candidates [33,39]. We assessed whether imputation accuracy would differ depending on the reference panel being composed of randomly selected individuals or older individuals and found no advantage in imputation accuracy when selecting a reference panel composed of older animals.

In addition to the observed increase in overall imputation accuracy, we found that increasing the size of the reference panel is especially efficient at increasing the individual imputation accuracy of SNP that exhibited below average imputation accuracy [20]. SNP with MAF below 0.1 were imputed poorly in comparison to SNP with MAF above 0.1 (accuracy measure  $R^2$ ,  $IA_{MAF}$ ), but as reference panel size increased imputation accuracy of these SNP improved, and for imputation based on 1024 haplotypes we observed a uniform distribution of imputation accuracy (quantified as  $R^2$ ) across levels of MAF (Additional file 1: Figure S1). An increase in the size of the reference panel increases the precision of estimated frequencies of haplotypes containing rare alleles, which appears to more efficiently boost imputation accuracy for the corresponding SNP [20]. SNP located in the 10% chromosomal extremes (5% on either side) also had on average lower imputation accuracy than the remaining SNP. As reference panel size was increased very little improvement could be observed in imputation accuracy of SNP located in the center of the chromosome, due to these SNP already being imputed with accuracy close to 1. However, imputation accuracy of SNP in the chromosome ends improved as reference haplotypes were added to the panel, until reaching accuracy within 0.02 points of the average imputation accuracy of SNP in the remainder

of the chromosome for imputation based on a reference panel containing 1024 haplotypes (Additional file 2: Figure S2).

Although, we did not assess the accuracy of GEBV prediction based on imputed genotypes in this paper, we can use results from dairy cattle breeding that show the promise of imputed genotypes to predict GEBV. Based on the average imputation accuracy we observed for Yorkshire pigs and previous results for GEBV prediction based on imputed genotypes in dairy cattle we could expect that losses in accuracy of GEBV prediction as a result of genotype imputation will be negligible. Wiggans et al. [40] and Dassonneville et al. [9] reported correlation of GEBV from imputed genotypes ( $IA \geq 0.96$ ) with GEBV estimated from high density genotypes larger than 0.93. Moreover, Weigel et al. [10], reported a loss in accuracy of GEBV, estimated as the correlation between GEBV and direct genomic value, between 0 and 5% when using genotypes imputed with low accuracy ( $IA = 0.91$ ). Since our estimates of imputation accuracy in the Yorkshire population are within the range of those reported in dairy cattle [9,10,33,40], we expect GEBV estimated from imputed genotypes in Yorkshire pigs to be as accurate as those currently used in the dairy breeding industry. Furthermore, these results are expected to hold in other swine breeds with similar levels of LD [23].

## Conclusion

In conclusion, high ( $IA \geq 0.95$ ) genotype imputation accuracy can be achieved in pigs combining the newly available commercial 9K tagSNP set and a relatively small reference haplotype panel (128 haplotypes), even when imputation is based only on population-wide LD. Further improvements in imputation accuracy could be achieved through the inclusion of additional reference animals ( $IA = 0.97$  with 512 reference haplotypes) and the use of pedigree relations between reference and imputation animals in the imputation algorithm [35,36,40]. An important result from this study is that an efficient design for reference panel construction is randomly sampling individuals instead of specifically sampling older animals or trios. In addition, a relatively small panel of reference haplotypes ( $\geq 128$ ) can efficiently serve as a reference panel for genotype imputation, such that any available high density genotypes in a livestock population could potentially serve this purpose. For the pig species such panels are already available for several populations [23]. Finally, prospects for the use of imputed genotypes in GEBV prediction are very positive based on the results from dairy breeding that routinely use similarly accurately imputed genotypes for genomic evaluation [9,10,40]. The methodology used in this paper for construction of tagSNP sets and reference haplotype panels can be easily applied in any future study population. Code and data to obtain and

reproduce the results presented is publicly available at [https://www.msu.edu/~steibelj/JP\\_files/imputation.html](https://www.msu.edu/~steibelj/JP_files/imputation.html).

## Additional files

**Additional file 1: Figure S1.** Effect of reference panel size on imputation accuracy of SNP as a function of their MAF. Weighted mean average imputation accuracy (quantified as  $R^2$ ) as a function of MAF depicted for imputation based on haplotype reference panels of increasing size.

**Additional file 2: Figure S2.** Effect of reference panel size on imputation accuracy of SNP as a function of scaled physical location. Weighted mean average imputation accuracy (quantified as  $I_A$ ) as a function of the scaled chromosomal location for imputation based on haplotype reference panels of increasing size.

## Authors' contributions

JPS, ROB, CS, and CWE designed the experiments. CS, ROB, and CWE identified samples for data collection and oversaw the collection of tissue samples. CWE oversaw DNA isolation, sample quality and commercial genotyping. CS, JF, and ROB provided pedigree information and individual animal identification. CPT selected the tagSNP for the commercial platform. JPS and YMB completed statistical analysis and wrote the manuscript. All authors read and approved the paper.

## Acknowledgements

This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2010-65205-20342 from the USDA National Institute of Food and Agriculture, and by funding from the National Swine Registry. Technical assistance of Nancy Raney and Whitney Hosier is appreciated. Computer resources and programming advice were provided by the Michigan State University High Performance Computing Center (HPCC).

## Author details

<sup>1</sup>Department of Animal Science, Michigan State University, East Lansing, MI, USA. <sup>2</sup>Department of Fisheries & Wildlife, Michigan State University, East Lansing, MI, USA. <sup>3</sup>The Maschhoffs, Carlyle, IL, USA. <sup>4</sup>National Swine Registry, West Lafayette, IN, USA. <sup>5</sup>Bovine Functional Genomics Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD, USA.

Received: 25 October 2012 Accepted: 29 January 2013

Published: 21 February 2013

## References

- Boichard D, Chung H, Dasonneville R, David X, Eggen A, Fritz S, Gietzen KJ, Hayes BJ, Lawley CT, Sonstegard TS, Van Tassell CP, Vanraden PM, Viaud-Martinez KA, Wiggins GR: **Design of a Bovine Low-Density SNP Array Optimized for Imputation.** *PLoS ONE* 2012, **7**(3):e34130. [http://dx.plos.org/10.1371/journal.pone.0034130].
- Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TPL, Sonstegard TS, Tassell CPV: **Development and Characterization of a High Density SNP Genotyping Assay for Cattle.** *PLoS One* 2009, **4**(4):e5350. [http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0005350].
- VanRaden PM, O'Connell JR, Wiggins GR, Weigel KA: **Genomic evaluations with many more genotypes.** *Genet Sel Evol* 2011, **43**:10. [http://www.ncbi.nlm.nih.gov/pubmed/21366914].
- Groenen MM, Wahlberg P, Foglio M, Cheng HH, Megens HJ, Crooijmans RPM, Besnier F, Lathrop M, Muir WM, Wong GKS, Gut I, Andersson L: **A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate.** *Genome Res* 2009, **19**(3):510–519. [http://www.ncbi.nlm.nih.gov/pubmed/19088305].
- Archibald AL, Cockett NE, Dalrymple BP, Faraut T, Kijas JW, Maddox JF, McEwan JC, Hutton Oddy V, Raadsma HW, Wade C, Wang J, Wang W, Xun X: **The sheep genome reference sequence: a work in progress.** *Animal Genet* 2010, **41**(5):449–53. [http://www.ncbi.nlm.nih.gov/pubmed/20809919].
- Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, Hansen MS, Hedegaard J, Hu ZL, Kerstens HH, Law AS, Megens HJ, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TPL, Schnabel RD, Tassell CPV, Taylor JF, Wiedmann RT, Schook LB, Groenen MAM: **Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology.** *PLoS One* 2009, **4**(8):e6524. [http://www.ncbi.nlm.nih.gov/pubmed/19654876].
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: **Invited review: Genomic selection in dairy cattle: Progress and challenges.** *J Dairy Sci* 2009, **92**(2):433–443. [http://www.ncbi.nlm.nih.gov/pubmed/19164653].
- Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**(4):1819–1829. [http://www.genetics.org/content/157/4/1819.abstract].
- Dasonneville R, Brondum RF, Druet T, Fritz S, Guillaume F, Guldbandsen B, Lund MS, Ducrocq V, Su G: **Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations.** *J Dairy Sci* 2011, **94**(7):3679–3686. [http://www.ncbi.nlm.nih.gov/pubmed/21700057].
- Weigel KA, de los Campos G, Vazquez AI, Rosa GJM, Gianola D, Tassell CPV: **Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle.** *J Dairy Sci* 2010, **93**(11):5423–5435. [http://www.ncbi.nlm.nih.gov/pubmed/20965358].
- Berry DP, Kearney JF: **Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection.** *Animal* 2011, **5**(08):1162–1169. [http://www.ncbi.nlm.nih.gov/pubmed/22440168].
- Habier D, Fernando RL, Dekkers JCM: **Genomic Selection Using Low-Density Marker Panels.** *Genetics* 2009, **182**:343–353. [http://www.genetics.org/content/182/1/343.abstract].
- Browning BL, Browning SR: **A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals.** *Am J Hum Genet* 2009, **84**(2):210–223. [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2668004/].
- Scheet P, Stephens M: **A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.** *Am J Hum Genet* 2006, **78**(4):629–644. [http://www.ncbi.nlm.nih.gov/pubmed/16532393].
- Hickey JM, Crossa J, Babu R, de los Campos G: **Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs.** *Crop Sci* 2012, **52**(2):654. [https://www.agronomy.org/publications/cs/abstracts/52/2/654].
- Hayes BJ, Bowman PJ, Daetwyler HD, Kijas JW, van der Werf JHJ: **Accuracy of genotype imputation in sheep breeds.** *Anim Genet* 2012, **43**:72–80. [http://www.ncbi.nlm.nih.gov/pubmed/22221027].
- He J, Zelikovsky A: **Informative SNP selection methods based on SNP prediction.** *IEEE Trans Nanobioscience* 2007, **6**:60–67. [http://www.ncbi.nlm.nih.gov/pubmed/17393851].
- Qin ZS, Gopalakrishnan S, Abecasis GR: **An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria.** *Bioinf* 2006, **22**(2):220–225. [http://www.ncbi.nlm.nih.gov/pubmed/16269414].
- He J, Zelikovsky A: **MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression.** *Bioinf* 2006, **22**(20):2558–2561. [http://bioinformatics.oxfordjournals.org/content/22/20/2558.full].
- Howie B, Marchini J: **Genotype Imputation with Thousands of Genomes.** *G3* 2011, **1**(457):13. [http://www.g3journal.org/content/1/6/457.full].
- Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P: **Genotype-imputation accuracy across worldwide human populations.** *Am J Hum Genet* 2009, **84**(2):235–250. [http://www.ncbi.nlm.nih.gov/pubmed/19215730].
- Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, Donnelly P, Consortium IH: **A comparison of phasing algorithms for trios and unrelated individuals.** *Am J Hum Genet* 2006, **78**(3):437–450. [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1380287/].

23. Badke YM, Bates RO, Ernst CW, Schwab C, Steibel JP: **Estimation of linkage disequilibrium in four US pig breeds.** *BMC Genomics* 2012, **13**:24. [http://www.biomedcentral.com/1471-2164/13/24].
24. de Roos aPW, Hayes BJ, Goddard ME: **Reliability of genomic predictions across multiple populations.** *Genetics* 2009, **183**(4):1545–1553. [http://www.genetics.org/content/183/4/1545.abstract].
25. Browning BL: **Documentation of BEAGLE 3.3.1.** [http://faculty.washington.edu/browning/beagle/beagle.html].
26. Zheng J, Li Y, Abecasis GR, Scheet P: **A comparison of approaches to account for uncertainty in analysis of imputed genotypes.** *Genet Epidemiol* 2011, **35**(2):102–110. [http://www.ncbi.nlm.nih.gov/pubmed/21254217].
27. Becker RA, Chambers JM, Wilks AR: *The new S language: A programming environment for data analysis and graphics* Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software; 1988 [http://adsabs.harvard.edu/abs/1988nsl.book....B].
28. Dassonneville R, Fritz S, Ducrocq V, Boichard D: **Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle.** *Journal of dairy science* 2012, **95**(7):4136–40. [http://www.ncbi.nlm.nih.gov/pubmed/22720970].
29. Cleveland WS, Grosse E, Shyu WM: **Local regression models.** *Statistical Models in S.* Edited by Chambers JM, Hastie TJ. Pacific Grove: Wadsworth & Brooks/Cole; 1992:309–376.
30. Steibel J, Wysocki M, Lunney J: **Assessment of the swine protein-annotated oligonucleotide microarray.** *Anim Genet* 2009, **40**(6):883–893. [http://www.ncbi.nlm.nih.gov/pubmed/19515086].
31. Abecasis GR, Cherny SS, Cookson WO, Cardon LR: **Merlin—rapid analysis of dense genetic maps using sparse gene flow trees.** *Nat Genet* 2002, **30**:97–101. [http://www.ncbi.nlm.nih.gov/pubmed/11731797].
32. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, Van Der Werf JH: **A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes.** *Genet Sel Evol GSE* 2011, **43**:12. [http://www.ncbi.nlm.nih.gov/pubmed/21388557].
33. Weigel KA, Tassell CPV, O'Connell JR, VanRaden PM, Wiggans GR: **Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms.** *J Dairy Sci* 2010, **93**(5):2229–2238. [http://www.ncbi.nlm.nih.gov/pubmed/20412938].
34. Vittinghoff E, Shiboski SC, Glidden DV, McGulloch CE: *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models.* New York: Springer; 2005.
35. Gualdrón Duarte JL, Bates RO, Ernst CW, Raney NE, Cantet RJC, Steibel JP: **Genotype imputation accuracy in an F2 pig cross using high density and low density SNP panels [abstract].** ADSA-ASAS Joint Annual Meeting: Phoenix; 2012:W76.
36. Huang Y, Hickey JM, Cleveland MA, Maltecca C: **Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost.** *Genet, Sel, Evol: GSE* 2012, **44**:25. [http://www.gsejournal.org/content/44/1/25].
37. Browning SR, Browning BL: **Haplotype phasing: existing methods and new developments.** *Nat Rev Genet* 2011, **12**(10):703–714. [http://www.ncbi.nlm.nih.gov/pubmed/21921926].
38. Welsh CS, Stewart TS, Schwab C, Blackburn HD: **Pedigree analysis of 5 swine breeds in the United States and the implications for genetic conservation.** *J Anim Sci* 2010, **88**(5):1610–1618. [http://www.ncbi.nlm.nih.gov/pubmed/20190174].
39. Zhang Z, Druet T: **Marker imputation with low-density marker panels in Dutch Holstein cattle.** *J Dairy Sci* 2010, **93**(11):5487–5494. [http://www.ncbi.nlm.nih.gov/pubmed/20965364].
40. Wiggans GR, Cooper TA, VanRaden PM, Olson KM, Tooker ME: **Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation.** *J Dairy Sci* 2012, **95**(3):1552–1558. [http://www.ncbi.nlm.nih.gov/pubmed/22365235].

doi:10.1186/1471-2156-14-8

Cite this article as: Steibel et al.: Methods of tagSNP selection and other variables affecting imputation accuracy in swine. *BMC Genetics* 2013 **14**:8.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

