How much extra information is gained using imputed genotype data?

S. Clark [1,2], N. Duijvesteijn [2] and J.H.J van der Werf [1,2]

[1] School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia
sam.clark@une.edu.au (Corresponding Author)
[2] Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW 2351, Australia

Summary

Genotype imputation has been discussed widely as a tool to increase the statistical power associated with genome wide association studies (GWAS) and genomic prediction. Previous studies have examined the performance of imputation by evaluating how well a validation set has been predicted. The aim of this study was to examine the amount of extra information added by utilising genotype imputation. The number of new haplotype combinations, between adjacent loci, was estimated for multiple genotype densities from an Australian Sheep dataset. In our example, using genotypes from OAR6, imputation increased the number of haplotypes for 81% of the regions when imputing from 12k to 50k. Large distances between adjacent low density markers resulted in higher numbers of new haplotypes. This also corresponded to a greater proportion of low frequency haplotypes. When imputing from HD to WGS no information was added for 13% of the regions. The number of new haplotypes was directly related to the number of animals in the reference dataset and the distance between adjacent markers. Estimating the number of new haplotypes from imputation provides an understanding about the value of imputation and can be utilised to help design of reference genotype datasets.

Keywords: imputation, genomic selection. GWAS, power, QTL detection, sheep, whole genome sequence, high density.

Introduction

Genomic selection is now common place in most livestock breeding programs. Over the past decade the number of genotypes used in genomic selection has increased significantly such that the use of sequence data, which includes all of the variants, is now a reality. Although possible, the cost associated with obtaining whole genome sequence on all individuals is still prohibitive. In most cases those utilising sequence data have sequenced a small number of animals and utilised genotype imputation, from lower marker densities, to expand the sample size to include all previously genotyped (and phenotyped) individuals (Van Binsbergen et al 2015).

Genotype imputation has been discussed widely as a tool to increase the power associated with genome wide association studies (GWAS) and genomic prediction. Many studies have shown that genotype imputation allows for the capture of most of the benefits of increasing marker density at a fraction of the cost (Marchini and Howie 2010). Much time and effort has been put into the development of computationally efficient imputation methods. Most recent imputation algorithms (HMM or Heuristic methods) centre on the concept that the haplotypes to be imputed are an assortment of those found in a reference dataset (Browning & Browning 2016). Essentially they utilise information from adjacent

markers to infer un-genotyped regions. Most imputation studies define the success of imputation using 2 major methods; the accuracy of imputation and concordance by various types of validation (Daetwyler et al 2011). Given the population structure associated with animal breeding programs, high values are frequently observed. It is often unclear whether these high values, of accuracy and concordance, are due to population structure or whether they are a representation of new information gained from the imputation process. For example it could be hypothesized that the new imputed haplotypes are just an extension of existing haplotypes and no new information is added.

The aim of this study was to examine how many new haplotypes are created due to the use of genotype imputation.

Material and methods

Assessing new haplotypes

A simple method to estimate the number of new haplotypes added between two adjacent loci was undertaken to observe how much added information was incorporated from the low density panel. Haplotypes which occurred fewer than 5 times were removed, as they coincided with likely imputation errors or very low frequency SNP.

Data used

To examine the number of new haplotypes added by imputation; a number of data sets from the Australian Sheep CRC were used. This data consisted of ~45,000 genotyped animals from multiple sheep breeds. All sheep were genotyped on either the Illumina 50k Ovine SNP chip; the ovine LD (12k) SNP chip (Bolormaa et al 2015); the Ovine Infinium® HD SNP BeadChip, or whole genome sequenced (WGS). Three imputation scenarios were tested; the first scenario (LD-50K) was to observe the impact of imputing 21,525 animals from 12k to 50k (reference 23830 animals) using beagle 3.3.2 (Browning & Browning 2011). The second scenario (50K-HD) was to observe the impact of imputing from 23830 animals from 50k to 600k (reference 2400 animals). The third scenario (HD-WGS) was to observe the impact of imputing 2400 animals from 600k to WGS (reference 781 animals) (Bolormaa et al 2018). All animals used in each scenario were selected based on the real genotype being present. Animals with imputed genotypes were not used in subsequent analysis of higher density imputation. Comparisons presented in this study refer to SNP on Chromosome 6 (OAR6).

Results and Discussion

Most regions had significantly more haplotypes than expected. Given each genotype was coded (0,1,2) there was a base expectation of 9 haplotypes (3^2) per region. In some cases fewer haplotypes were observed. Imputation increased the number of haplotypes for 81% of the regions when imputing from 12k to 50k (Table 1). When imputing from 50k to HD, 82% of the regions had more than the expected number of haplotypes. Similarly, when imputing from HD to sequence 87% of the regions had more than the expected number of haplotypes.

Table 1. Summary of the number of new haplotypes for each imputation scenario

Summary of new haplotypes					
Test	Mean	Min	Max	[2]n>9	n SNP regions
LD-50K	201	6	960	460	563
50K-HD	86	2	350	2000	2410
HD-WGS[1]	45	1	121	21959	24417

[1]WGS included 1,565,530 SNP on OAR 6
[2]Number of regions with greater than 9 haplotypes

Large distances between adjacent low density markers resulted in higher numbers of new haplotypes. This also corresponded to a greater proportion of low frequency haplotypes. For example Figure 1 shows a region on OAR 6 that corresponds to a known QTL associated with body weight (Al-Mamun et al., 2015). This region (36.3 Mb to 37.2 Mb) is sparsely covered by the 12k chip and had 3300 unique haplotypes observed in that area. Only 19% (620) of the haplotypes were observed more than once. The initial value may seem high however there are ~14 million possible haplotype combinations for a 15 SNP haplotype, as observed in this region.

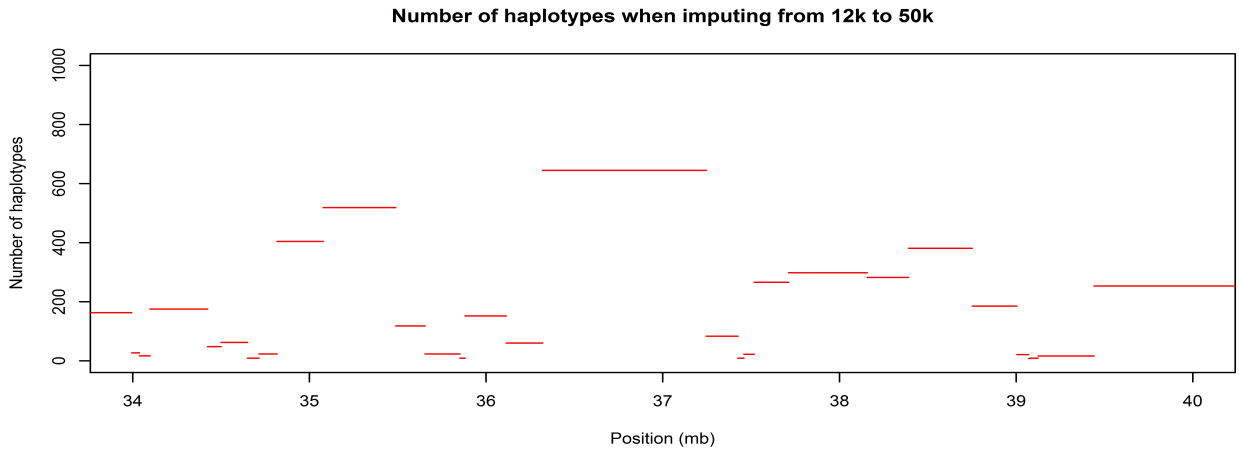

Figure 1. Number of new haplotypes observed in the OAR6 QTL region. QTL region associated with body weight

There were fewer new haplotypes per region as the density of the chip increased, however the haplotype size decreased substantially. As shown by Figure 2, the frequency of the number of new haplotypes changed considerably when comparing 12k-50k and HD-WGS. The major factor that affected the number of haplotypes observed were the number of animals in the reference population (23830 vs 781) and the distance between existing markers.

Conclusion

Estimating the number of new haplotypes from imputation provides an understanding about the value of imputation and can be utilised to help design of reference genotype datasets. To

get more value out of sequence information, and for that matter HD genotyping, more animals need to be genotyped with these higher density panels to increase the size of the reference dataset. To make the most out of this effort, algorithms that optimise which animals to sequence (or genotype) based on the probability of new haplotypes or that increase the frequency of rare haplotypes will be very useful (i.e. Gonen et al 2017).

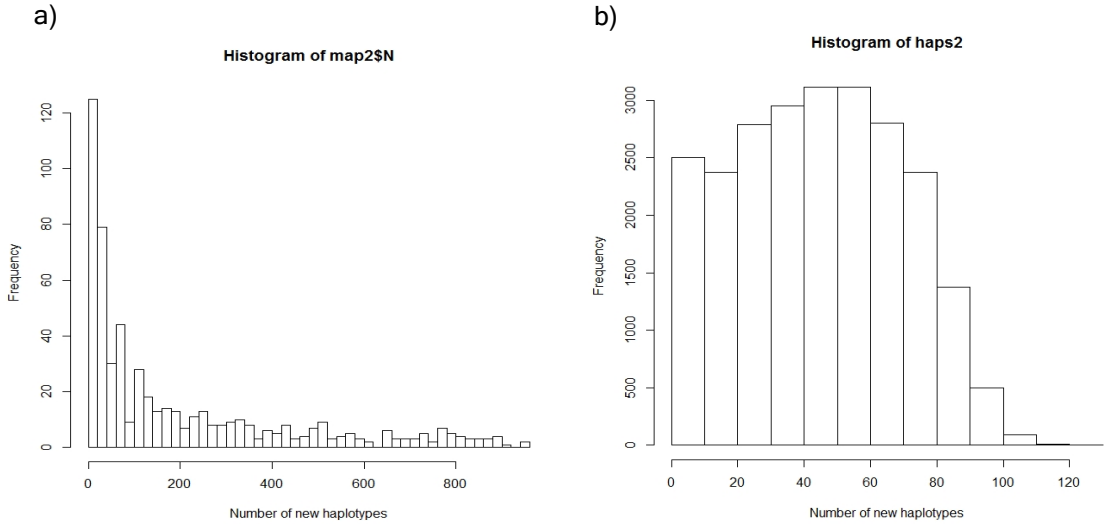

Figure 2. Frequency of different numbers of new haplotypes a) 12k-50k b) HD-WGS

List of References

Browning, B., Browning, S.R., 2016. Genotype Imputation with Millions of Reference Samples. AJHG, 98 (1): 116-126.

Bolormaa, S., Gore, K., van der Werf, J. H. J., Hayes, B. J. and Daetwyler, H. D., 2015. Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. Anim Genet, 46: 544–556.

Bolormaa, S., A. Chamberlain, J. H. J. van der Werf, H. D. Daetwyler & I. M. MacLeod, (2018). Evaluating the accuracy of imputed whole genome sequence in sheep. Proc. 10th World Congr. Genet. Appl. Livest. Prod

Gonen, S., Ros-Freixedes, R., Battagin, M., Gorjanc, G., and Hickey, J. M., 2017. "A method for the allocation of sequencing resources in genotyped livestock populations." GSE Evolution 49(1): 47.

Van Binsbergen, R., Calus, M. P. L., Bink, M. C. A. M., van Eeuwijk, F. A., Schrooten, C., & Veerkamp, R. F. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. GSE, 47(1): 71.

Marchini, J. and B. Howie 2010. "Genotype imputation for genome-wide association studies." Nat Rev Genet 11(7): 499-511.

Daetwyler, H. D., Wiggans, G. R., Hayes, B. J., Woolliams, J. A., and Goddard, M. E., 2011. "Imputation of Missing Genotypes From Sparse to High Density Using Long-Range Phasing." Genetics 189(1): 317.

Druet, T., Macleod, I. M., & Hayes, B. J. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity, 112(1): 39–47.

Al-Mamun HA, Kwan P, Clark SA, Ferdosi MH, Tellam R, Gondro C., 2015. Genome-wide association study of body weight in Australian Merino sheep reveals an orthologous region

on OAR6 to human and bovine genomic regions affecting height and weight. Genet Sel Evol. 47(1):66.