WHY CAN WE IMPUTE SOME RARE SEQUENCE VARIANTS AND NOT OTHERS?

B.J. Hayes[1,2,3], P.J. Bowman[1,3], H.D. Daetwyler[1,2,3] and M.E. Goddard[1,3,4]

[1]AgriBio, Centre for AgriBioscience, Biosciences Research, DEDJTR, Victoria, Australia
[2]Biosciences Research Centre, La Trobe University, Victoria, Australia
[3]Dairy Futures Cooperative Research Centre, Victoria, Australia
[4]Faculty of Veterinary & Agricultural Sciences, University of Melbourne, Victoria, Australia

SUMMARY

We investigated how well rare variants can be imputed, using 1000 bull genomes sequence data set (1147 sequences) as a reference for imputation, and a target set of dairy cattle with 630K SNP genotypes, that were also genotyped for four rare recessive defects (BLAD, CVM, HH1 and JH1). The proportion of carriers correctly imputed ranged from 1, for JH1, to 0.04 for CVM. There was a general trend for the proportion of carriers correctly imputed to increase as the frequency of the rare allele increased. CVM did not follow this trend – the frequency of the rare allele for this locus was 10 times higher than for BLAD, but proportion of carriers correctly imputed was much lower than BLAD. On closer inspection, the core haplotype of sequence variants common to all CVM carriers was found in many non-carriers, and even in breeds other than Holstein (the disease has only been reported in Holstein). This was in contrast to JH1, where the core haplotype shared by carriers was unique to carriers, and was not found in other breeds. These results shed light on why we can impute some rare sequence variants well, while others are very difficult to impute.

INTRODUCTION

One motivation for using whole genome sequence data in genomic prediction and genome wide association studies (GWAS) is that whole genome sequence data will include rare variants which may explain some variation in the targeted complex traits. SNP arrays have limited power to capture this variation, as the SNP on these arrays are selected to have high minor allele frequency (MAF), and are therefore unlikely to be in high linkage disequilibrium with the rare variants. The cost of whole genome sequencing is currently too high to sequence the large number of individuals required for accurate genomic predictions or powerful GWAS. Therefore an alternative strategy has been proposed – sequence a proportion of the individuals in the population (1000 Genomes Project Consortium *et al*. 2012), or preferably the key ancestors of the population (eg Daetwyler *et al*. 2014), and then impute the sequence variants into all individuals genotyped with SNP arrays. How much variation is explained by rare variants in subsequent genomic predictions or GWAS will then depend on how much the rare variants truly explain, and the accuracy of imputing these rare variants.

Here we investigate how well rare variants can be imputed, using 1000 bull genomes sequence data set as a reference, and a target set of dairy cattle that were actually genotyped for four rare recessive defects. In order to gain insights into parameters affecting accuracy of imputation of rare variants, we investigated the length of core haplotype surrounding the disease allele for each recessive defect, the occurrence of this haplotype (minus the disease allele) in non-carriers and the frequency of this haplotype in breeds other than the one in which the disease occurs.

MATERIALS AND METHODS

Carrier status (from genotyping the causal mutation was available for four recessive diseases - Bovine leukocyte adhesion deficiency (BLAD, Shuster *et al*. 1992), complex vertebral malformation (CVM, Thomsen *et al*. 2006), Holstein Haplotype 1 (HH1, Adams *et al*. 2012) and

Jersey Haplotype 1 (JH1, Sonstegard *et al*. 2013). Genotypes for these mutations were available for 5987 Holstein (BLAD, CVM), 707 Holstein (HH1) and 16 Jersey bulls (JH1), respectively, as well 630K Bovine HD real or imputed SNP genotypes (eg. Erbe *et al*. 2012). In order to impute the BLAD, CVM, HH1 or JH1 genotypes into these animals, to compare with their actual genotypes, we used a reference data set of 1147 bulls and cows of 20 breeds with whole genome sequence, These reference animals were sequenced at between 4 and 40 times coverage, with an average of 11.2x, from 1000 bull genomes Run4.0. The breeds with largest number of sequenced individuals were Holstein, Angus and Fleckvieh. Variant calling and filtering was as described by Daetwyler *et al*. (2014). Variants with less than 4 copies of the minor allele were removed. We checked that all known carriers of BLAD, CVM, HH1 or JH1 that had whole genome sequence data (eg were part of the 1000 bull genomes) were genotyped correctly for these mutations, this was the case. Two imputation strategies to impute sequence variants into the target populations were tested, Fimpute (Sargolzaei *et al*. 2014) or Beagle phasing followed by Minimac imputation (Howie *et al*. 2012). Differences between these programs are that Fimpute uses full pedigree information, while Minimac does not, and Fimpute considers variable length haplotypes, starting from long haplotypes, when deciding if a pair of animals share a haplotype. Actual genotypes of the recessive lethals for target animals were not included when target animals were imputed to whole genome sequence genotypes. Imputed genotypes were then compared to actual genotypes for these defects.

RESULTS AND DISCUSSION

The proportion of genotypes imputed correctly was close to one for all loci, Table 1.

Table 1. Proportion of genotypes and proportion of carriers correctly imputed for four genetic defects.

	BLAD	CVM	HH1	JH1
Chromosome	1	3	5	15
Location (bp)	145114963	43412427	63150400	15707169
Frequency	0.001	0.010	0.025	0.156
Bulls genotyped in target population	5987	5987	707	16
Genotypes imputed correctly				
Fimpute	5970	5836	701	16
Minimac	5860	5860	705	16
Prop. genotypes imputed correctly				
Fimpute	0.997	0.97	0.99	1.00
Minimac	0.98	0.98	0.997	1.00
Number of carriers	17	123	35	5
Carriers correctly imputed				
Fimpute	13	5	29	5
Minimac	11	12	33	5
Prop. carriers correctly imputed				
Fimpute	0.77	0.04	0.83	1.00
Minimac	0.65	0.10	0.94	1.00

However this is a poor measure of how well imputation has performed for rare variants, given the high probability of filling in the correct genotype by chance (a very high proportion of animals are homozygous for the non-disease allele).

A better measure of how well imputation has performed is the proportion of carriers correctly imputed – for GWAS and genomic prediction, this will determine how well the SNP effect can be estimated. This ranged from 1, for JH1, to 0.04 for CVM. There was a general trend for the proportion of carriers correctly imputed to increase as the frequency of the rare allele increased. The imputation of CVM genotypes did not follow this trend – the frequency of the rare allele for this locus was 10 times higher than for BLAD, but the proportion of carriers correctly imputed was much lower than for BLAD.

To investigate why this might be the case, and given imputation is based on haplotype information shared between individuals, we determined the length of haplotype in the sequenced bulls (from the 1000 bull genomes project) surrounding the rare allele of each locus that was common between all carriers, the "core haplotype". To do this, we allowed for sequencing error, such that the shared haplotype was considered to end only when there were at least two differences in the alleles of the haplotype of the carriers (eg one difference was considered to be likely sequencing error - in fact there were only one or at most two instances of this per disease). HH1 had the longest core haplotype, while CVM had the shortest, Table 2. We then investigated how many non-carriers amongst all the Holstein sequenced bulls (for BLAD, CVM, and HH1) or Jersey sequenced bulls (JH1) had the core haplotype (not considering the disease allele itself). This ranged from zero, for JH1, to 159, for CVM. For all diseases except JH1, the core haplotype also occurred in other breeds (where these diseases have never been observed), though at very low frequency, and in only a small number, except for CVM.

Table 2. Length of core haplotype shared by all whole genome sequenced carriers of the disease (rare) allele for four lethal recessive diseases, number of non-carriers in which core haplotype is found, and number of other breeds in which core haplotype is found.

	BLAD	CVM	HH1	JH1
Number of carriers with whole genome sequence	6	30	7	12
Variants in core haplotype (shared by carriers)	302	93	437	633
Length of core haplotype (bp)*	40,362	21,020	57,173	48,608
Number of non-carriers in which core haplotype is found	4	159	1	0
Number of other breeds in which core haplotype is found	1	24	2*	0

*One of these was Danish red, which has Holstein introgressions

Given these results, we can start to speculate why the imputation of CVM genotypes is so poor, while for JH1, HH1 and BLAD imputation is more precise. The background haplotype in which the CVM mutation occurs, appears to be very common, even across breeds. It is likely that the CVM mutation occurred recently into this common haplotype background, such that there are otherwise identical haplotypes at reasonable frequency, without the mutation. This makes imputation, which is based on haplotype information, very challenging. In contrast, the JH1 mutation is imbedded in a longer haplotype which was likely at a lower frequency at the time the mutation occurred, such that carriers of the haplotype are also very likely to be carriers of the mutation as well. Parameters such as the frequency of the core haplotype into which the rare mutation occurred likely explain results from other studies as well, such as those of Bouwman *et al*. (2014), where reference sets for imputation which included multiple breeds improved accuracy of imputing a proportion of rare variants, but not others, compared to single breed reference sets.

Is there any way to improve the precision of imputing rare variants in light of the above? One of the first tasks is to reduce the error rate of genotyping variants from the whole genome sequence data – this complicates the identification of the core haplotype shared by carriers of the rare allele, and importantly might reduce the length of the core haplotype that can be confidently identified, which will reduce the accuracy of imputation (longer shared haplotypes between individuals lead to more precise imputation, eg Sargolzaei *et al*. 2014). Phasing errors are also important (phasing is necessary for imputation both in the sequenced animals and in the animals genotyped with 630K, and there could be errors in either), and are compounded by genotyping errors. So reducing genotyping errors could also improve the accuracy of phasing the data, which is desirable as any switch errors (false positive recombinations), if these are in the reference animals, will also reduce precision of imputation in the target animals. A practical way to remove some genotyping errors would be to run imputation for very rare variants within a breed, or combine LD information across closely related breeds (based on Fst for example)), only considering variants that segregate within the breed or group of breeds. This would reduce the number of variants (per breed), and therefore the opportunities for genotyping error, by 50% (Daetwyler *et al*. 2014). Information could then be accumulated across breeds.

ACKNOWLEDGEMENTS

The authors thank all members of the 1000 Bull Genomes Consortium for provision of data.

REFERENCES

Chadeau-Hyam M., Hoggart C., O'Reilly P., Whittaker J., De Iorio M. and Balding D. (2008) *BMC Bioinformatics* **9**: 364.

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., Mason B.A. and Goddard M.E. (2012) *J. Dairy Sci.* **95**: 4114.

Shuster D.E., Kehrli M.E. Jr, Ackermann M.R., Gilbert R.O. (1992) *Proc Natl Acad Sci U S A.* **89**:9225.

Thomsen B., Horn P., Panitz F., Bendixen E., Petersen A.H., Holm L.E., Nielsen V.H., Agerholm J.S., Arnbjerg J., Bendixen C. (2006) . *Genome Res*. **16**:97.

Adams H.A., Sonstegard T., VanRaden P.M., Null D.J., Van Tassell C., *et al*.. (2012) *Plant and Animal Genome Meeting*, Poster P0555, June 14–18 2012, San Diego.

Sonstegard T.S., Cole J.B., VanRaden P.M., Van Tassell C.P., Null D.J., *et al*. (2013) *PLoS ONE* **8**: e54872.

Sargolzaei M., Chesnais J.P., Schenkel F.S. (2014) *BMC Genomics*. **15**:478.

Howie B., Fuchsberger C., Stephens M., Marchini J., Abecasis G.R. (2012) *Nat Genet* **44**:955

Daetwyler H.D., Capitan A., Pausch H., *et al*. (2014) *Nat Genet*. **46**:858.

Bouwman A.C., Veerkamp R.F. (2014) *BMC Genet.* **15**:105.

1000 Genomes Project Consortium, Abecasis G.R., Auton A., *et al*. (2012) *Nature.* **491**:56.