Contributed paper

GENOTYPING DAIRY CATTLE WITH SKIM-WHOLE-GENOME SEQUENCING AND IMPUTATION

H.D. Daetwyler[1,2], J. Li[3], C.J. Vander Jagt[1], I.M. MacLeod[1], J. Pickrell[3], M. Vasquez[3], J.Hoff[3] and A.J. Chamberlain[1]

[1] Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, VIC, 3083 Australia
[2] School of Applied Systems Biology, La Trobe University, Bundoora, VIC, 3083 Australia
[3] Gencove, New York, NY, 10016, USA

SUMMARY

Cost effective genotyping tools are essential for wide-spread use of genomics in research and industry. While the majority of large-scale industry implementations of genomic selection have relied on single nucleotide polymorphism (SNP) arrays, genotyping using skim-whole-genome sequencing (SWGS) is becoming more accurate and, due to large reductions in sequencing cost, SWGS genotyping is becoming price competitive with SNP arrays. In SWGS genotyping, a sample is sequenced to 0.5 or 1x read depth and imputed to full WGS with a reference population sequenced at higher read depth (e.g. 1000 Bull Genomes Project). Imputation software, such as Beagle, can directly impute SNPs from SWGS to high fold coverage WGS, but they were not designed to do so. Gencove has developed an imputation algorithm especially for this task, 1oimpute. We compared the genotyping and imputation accuracy of Beagle4.0 and 1oimpute in a sample of 31 Holstein, 55 Jersey, and 39 Jersey-Holstein crosses. Animals were sequenced to approximately 10-fold coverage and variants and genotypes were identified as part of 1000 Bull Genomes Run8. Each animal's sequence was then randomly down-sampled to 0.5 and 1-fold coverage, aligned to the reference assembly, and imputed either with Beagle4.0 or with 1oimpute. Imputed genotypes were compared to observed full-sequence genotypes via correlation and proportion correct (concordance). The mean per marker genotype correlation of the 16 million imputed SNP across all breeds was 0.78 (0.5x) and 0.84 (1x) for Beagle and 0.92 (0.5x) and 0.93 (1x) for 1oimpute. While the Beagle pipeline could be likely further improved, the results demonstrate that a purpose-built imputation method is required to perform accurate SWGS genotyping. The method is attractive as it can provide sequence density genotypes at a cost price point comparable to low or medium-density SNP arrays.

INTRODUCTION

The large scale implementation of genomic breeding approaches in industry (e.g. genomic selection) requires genotyping tools that are accurate and cheap. The lower the cost of genotyping, the more widespread the adoption of genomic selection. Therefore, the continued development and refinement of genotyping methods is crucial to realising genetic gain from genomics.

Whole-genome sequencing has always underpinned genotyping platform development through the discovery of genetic marker diversity, such as single nucleotide polymorphisms (SNP), from which a subset of markers can be chosen for routine genotyping. Whole-genome sequencing requires the preparation of a library that cuts DNA into segments (i.e. sequence reads) and attaches a barcode to each segment. Once barcoded, samples can be mixed and sequenced together and the data for each sample can be separated afterwards. This multi-plexing approach coupled with vastly increased sequence output of recent technologies are the primary reasons for the large reduction in sequencing costs over time. The amount of sequencing per position of the genome is called read depth (e.g. read depths of 8 to 20x are common in livestock populations).

The most widely used genotyping method in large livestock populations have been SNP chips, which are microarrays that can provide genotypes on a few to many thousands of SNP. SNP chips

are generally highly accurate, amenable to high-throughput methods, and deliver near complete data at the loci queried. Low to medium density SNP chips with approximately <10,000 and 50,000 markers are currently available at prices that warrant wide-spread use when compared to impact on farm profitability (e.g. Newton *et al.* 2018). Nevertheless, decreasing genotyping costs further would no doubt increase the use of genomic selection.

Another way to genotype individuals is through genome sequencing directly. The reduced cost of sequencing now makes routine genotyping with whole-genome sequence feasible when sequence depth per sample is kept to 1x read depth or less, so-called skim whole-genome sequencing (SWGS). Due to the low read depth, there are relatively few loci with enough reads to call genotypes accurately and the set of loci called differs for each individual in a population. SWGS could be improved by imputing missing genotypes and improving genotype accuracy of loci with insufficient reads. Several imputation programs are available, such as Beagle, Minimac3, and FImpute, but most have not been developed specifically for imputing SWGS. Gencove have developed an imputation algorithm (1oimpute) for SWGS adapting an methods by Li and Stephens (2003) to routinely impute SWGS genotype data.

Here we present a comparison of SWGS genotyping using 1oimpute and Beagle4.0 imputation in three dairy cattle breed groups, Holstein, Jersey and Holstein-Jersey crossbreds, sequenced at 0.5 and 1x read depth.

MATERIALS AND METHODS

Whole-genome sequencing and processing. Thirty-one Holstein, 55 Jersey, and 39 Holstein-Jersey crossbred bulls were whole-genome sequenced to an average depth of 10x. Raw sequence fastq data were provided to Gencove and each animal's sequences were downsampled to 0.5 and 1x read depth. Full, 0.5 and 1x sequences were quality controlled and aligned with BWA to the ARS-UCD-1.2 reference assembly (Rosen *et al.* 2020) to produce binary alignment (bam) files. Full sequences were included in Run8 of the 1000 Bull Genomes Project (Hayes & Daetwyler 2019) and processed as described in Daetwyler *et al*. (2017).

Genotype calling and imputation. Two parallel pipelines were implemented by Gencove and Agriculture Victoria (AgVic) for a total of four scenarios: Gencove 1oimpute 0.5 and 1x read depth and AgVic Beagle at 0.5 and 1x read depth.

Gencove used their imputation software 1oimpute, which implements the Li and Stephens model for a set of reads in each animal's bam file and a known set of phased variants in a reference panel (Li & Stephens 2003). The diploid genotype probabilites are estimated using a Hidden Markov Model (Wasik *et al.* 2019). Gencove used a multi-breed reference panel of 946 animals (including 184 Holstein and 15 Jersey) for each breed (Snelling *et al.* 2020). AgVic performed variant calling on SWGS bam files using GATK3.8 according to the 1000 Bull Genomes Project guidelines. The 1000 Bull Genomes Project Run8 multi-breed taurus dataset with 4109 animals (including 1200 Holstein and 120 Jersey) was used as the AgVic reference for imputation. Random missing genotypes in the reference set were imputed with Beagle4.0 (Browning & Browning 2009) and filtered to only include biallelic SNP whose alleles occur at least 4 times. SWGS genotypes were then imputed with Beagle4.0 utilising genotype probabilities (Browning *et al.* 2018), and imputed animals were removed from reference sets.

Imputation accuracy evaluation. The accuracy of imputation was calculated as the Pearson correlation and concordance of imputed SWGS genotypes (coded as 0, 1, 2) from each respective pipeline and raw full sequence genotypes from the 1000 Bull Genomes Run8. Concordance was calculated as the proportion of imputed genotypes matching full sequence genotypes. Further, these statistics were summarised in minor allele frequency (MAF) bins of 0.0-0.03, 0.03-0.06, 0.06-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5. Comparisons were restricted to the set of SNP imputed by both 1oimpute and Beagle5.1 and passing the GATK quality tranche threshold of 99.9.

RESULTS AND DISCUSSION

The SWGS process led to approximately 1.6 million SNP. This is substantially more than a 50,000 marker SNP chip, but the SWGS SNP would be called with lower accuracy. The number of SNP imputed across all bovine autosomes by both 1oimpute and Beagle was 16,488,621 and the set of overlapping loci between the two pipelines were >95%.

Table 1. Mean correlation and concordance per SNP of imputed and observed genotypes in Holstein (HOL), Jersey (JER) and Holstein-Jersey crossbreds (HOLJER) from 1oimpute (G) and Beagle (B) pipelines.

Read Depth	0.5x Read Depth						1x Read Depth					
Breed	HOL		JER		HOLJER		HOL		JER		HOLJER	
Method	G	B	G	B	G	B	G	B	G	B	G	B
Correlation	0.95	0.79	0.90	0.78	0.90	0.76	0.95	0.84	0.91	0.84	0.92	0.83
Concordance	0.98	0.88	0.96	0.89	0.96	0.87	0.98	0.91	0.97	0.92	0.96	0.91

SD across autosomes ~0.01

The 1oimpute pipeline achieved substantially higher mean correlations between imputed and observed genotypes across all 16 million SNP tested, with a difference of ~0.15 (Table 1). This trend was also observed when using concordance as the evaluation measure, though the advantage of 1oimpute over Beagle was slightly less at ~0.1 (Table 1). This is quite a marked improvement that would surely result in improved downstream analyses. Imputation performance was quite similar across the three breeds for both piplelines. Interestingly, 1oimpute managed to still outperform Beagle even though Beagle had approximately 7 times the number of Holstein and Jersey animals in its reference. We did also test Beagle5.1, but it performed very poorly (correlation reduced by ~0.2) as it does not utilise genotype probabilities. Slightly better imputation was observed when animals were sequenced at 1x versus 0.5x, although the difference was small, and suggests that 0.5x is likely sufficient for the 1oimpute algorithm. Both imputation methods provide metrics per SNP on their confidence in genotype accuracy, which can be used to filter data further.

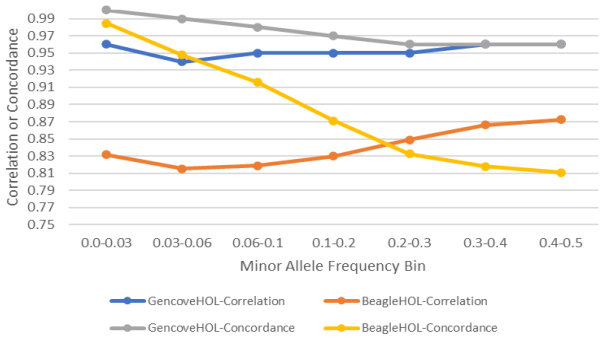

Figure 1. Mean correlation and concordance in minor allele frequency bins for Gencove 1oimpute and Beagle imputation for Holstein bulls with 1x sequence read depth.

It is well known that conventional imputation algorithm performance is substantially reduced for alleles with low frequency in the population (e.g. van Binsbergen *et al.* 2014). This was

confirmed for Beagle, where the correlation between imputed and observed genotypes in Holsteins was ~0.83 for loci with MAF < 0.03 (Figure 1). The reverse was observed for Beagle concordance, which was highest for the same low MAF bin. This occurs solely because most of the time, the most likely genotype will be correct and demonstrates the weakness of concordance as a measure of imputation accuracy, especially for low MAF SNP. In contrast, 1oimpute imputation correlations and concordance were consistently high (~0.95) across all MAF bins. Due to the high level of accuracy achieved by 1oimpute, both correlations and concordance were higher than Beagle across all MAF, though concordance did reach near 1.00 for low MAF SNP, indicating a small bias in this measure also for 1oimpute. Correlations and concordance followed similar levels and patterns across MAF for Jersey and crosses (data not shown).

The Beagle pipeline was not built specifically for imputing SWGS data with high proportion of missing genotypes and called genotypes with high uncertainty with different SNP called for each animal. Further improvement may be possible by filtering the SWGS genotypes for loci with read depth >5x. While this would further increase the proportion missing, it would provide more certain SNP genotypes to initiate the Beagle Hidden Markov Model. However, it seems unlikely that Beagle could achieve similar performance to 1oimpute even with these improvements. Recently, a new SWGS imputation method (GLIMPSE) has been published (Rubinacci *et al.* 2021), which seems competitive in accuracy with 1oimpute and testing with this method is underway.

The 1oimpute pipeline produces accurate genotypes at millions of loci and seems to overcome a traditional imputation bottleneck of accurately imputing lower MAF SNP. Industry application with the specific loci currently available on most SNP chips is therefore feasible, and for research applications, it is particularly useful to have access to many accurate genotypes across the MAF spectrum.

CONCLUSIONS

Substantially higher imputation accuracy was observed with Ioimpute than with Beagle. While the Beagle pipeline could be likely further improved, the results demonstrate that a purpose-built imputation method is required to perform accurate SWGS genotyping. The 1oimpute SWGS method is attractive as it can provide sequence density genotypes at a cost price point comparable to low or medium-density SNP chips.

ACKNOWLEDGEMENTS

The authors thank DairyBio, a joint venture project between Agriculture Victoria, Dairy Australia and The Gardiner Foundation, for funding and the 1000 Bull Genomes Project for use of whole-genome sequence data.

REFERENCES

Browning B.L. & Browning S.R. (2009) *Am. J. Hum. Genet.* **84**, 210-23.
Browning B.L., Zhou Y. & Browning S.R. (2018) *Am. J. Hum. Genet.* **103**, 338-48.
Daetwyler H.D., Brauning R., …, Kijas J.W. (2017) In: *Proc of AAABG*, Townsville, AUS.
Hayes B.J. & Daetwyler H.D. (2019) *Ann. Rev. Anim. Biosci.* **7**, null.
Li N. & Stephens M. (2003) *Genetics* **165**, 2213 - 33.
Newton J.E., Hayes B.J. & Pryce J.E. (2018) *J. Dairy Sci.* **101**, 6159-73.
Rosen B.D., Bickhart D.M., …, Medrano J.F. (2020) *GigaSci.* **9**, giaa021.
Rubinacci S., Ribeiro D.M., …, Delaneau O. (2021) *Nat Genet* **53**, 120-6.
Snelling W.M., Hoff J.L., …, Pickrell J.K. (2020) *Genes (Basel)* **11**.
van Binsbergen R., Bink M.C., …, Veerkamp R.F. (2014) *Genet. Sel. Evol.* **46**, 1-13.
Wasik K., Berisa T., …, Cox C. (2019) *bioRxiv*, 632141.