Contributed paper

CURRENT CHALLENGES FOR IMPUTATION OF SNP CHIPS TO WHOLE- GENOME SEQUENCE IN CATTLE AND SHEEP

I.M. MacLeod[1], S. Bolormaa[1], C.J. Vander Jagt[1], T.V. Nguyen[1], A.J. Chamberlain[1] and H.D. Daetwyler[1,2]

[1] Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, VIC, 3083 Australia
[2] School of Applied Systems Biology, La Trobe University, Bundoora, VIC, 3083 Australia

SUMMARY

Imputation to whole-genome sequence data has been successfully exploited in livestock for fine-mapping causal variants, meta-GWAS and increasing the accuracy of genomic prediction. However, imputation of sequence variants from marker panel (SNP chip) genotypes involves several key challenges that do not generally cause issues for SNP chip level imputation. Here we consider the challenges and potential solutions for issues such as rare variants, sequencing errors, misalignment in regions with large segmental duplications and/or copy number variants.

INTRODUCTION

Imputation of genotypes to sequence generally requires that target animals first have imputed or real marker panel (SNP chip) genotypes. Then the missing sequence variants between the markers are filled in using a reference set of real sequence genotypes. Imputation algorithms rely on the premise that animals sampled from a population will share a mosaic of haplotypes along the chromosome in common with one or more animals in the population. Even across breeds there are shared haplotypes due to their common ancestral origins. The observed length of the shared haplotypes depends on the marker density, local recombination rates, effective population size and importantly the level of relationships between the target individuals and the reference set. In livestock, it is commonplace to impute genotypes from lower density SNP chips to higher density chips. This imputation is highly accurate using a range of software (Calus *et al.* 2014) and has enabled genomic prediction of breeding values to become routine in the dairy, beef and sheep industries.

Imputation to whole-genome sequence from SNP panel genotypes is routinely undertaken for livestock research. The use of imputed sequence has been demonstrated to enable fine mapping of causal variants (e.g. Pausch *et al.* 2017), to facilitate meta Genome-Wide Association Studies (e.g. Bouwman *et al.* 2018) as well as increasing the accuracy of genomic prediction (e.g. Brøndum *et al.* 2015; Moghaddar *et al.* 2019; Xiang *et al.* 2021).

However, huge challenges remain compared to SNP chip level imputation for several reasons. First, 99% of the sequence variants are missing in high density SNP chip genotypes (HD: ~600k SNP) and the reference sequence data has higher error rates than SNP chip genotypes. This affects the accuracy of determining matching haplotypes between target and reference animals. Second, a large proportion of the sequence variants are less common (Minor Allele Frequency, MAF < 0.01) or rare compared to those selected for industry SNP chips and therefore may not be in strong linkage disequilibrium (LD) with the more common SNP on chips. This leads to inaccuracies for matching target to reference haplotypes. Third, it is costly to develop and maintain large representative sequence reference sets: a task that in addition to sequencing, requires considerable computational resources. Therefore, an attractive solution is for research groups to continue global collaborations to ensure that the databases continue to develop and grow by sharing costs/resources for sequence processing, storage and access.

The aim of this paper is to use examples from our own imputed and real sequence data to demonstrate the impact of some of the above challenges and briefly discuss potential solutions.

MATERIALS AND METHODS

We imputed sequence data into over 46,000 sheep and over 200,000 cattle using Minimac3 and pre-phased with Eagle software following Pausch *et al.* (2017). The sheep in the target set represented a range of breeds and crosses common to the Australian sheep industry, while the target cattle were dairy breeds and their crosses (mainly Holstein, Jersey and Australian Reds). Both sheep and cattle target populations had been imputed first to HD genotypes (~600k SNP). The sheep sequence reference used for imputation included 726 animals from European breeds and crosses in SheepGenomesDB Run2 (Daetwyler *et al.* 2017). The reference cattle sequences were from *Bos taurus* Run 6 and Run 7 of the 1000 Bull Genomes project (Hayes & Daetwyler 2019) and included 2333 and 3090 animals representing > 50 breeds and crosses. There were several key differences in the Run 6 (Daetwyler *et al.* 2017) and Run 7 pipeline: Run 6 was aligned to the UMD3.1 reference genome, while Run 7 used the improved ARS-UCD1.2 reference genome (Rosen *et al.* 2020). Run7 used GATK v3.8 for variant calling instead of Samtools (Run 6).

Prior to imputation, the variants called in the sheep and cattle reference sequences were pre-filtered to retain only bi-allelic variants (most imputation algorithms do not impute multiallelic variants) with minor allele counts of 4 or more (to remove variants that may be sequencing errors or so rare they cannot be well imputed). Additional pre-filtering was applied in Run 7 where we retained variants with Beagle R2 >0.9 (from the imputation of missing genotypes) and variants in GATK Tranche 99.0 or better. We also identified chromosome segments of ≥ 0.5Mb with excessive heterozygosity among genotyped individuals: i.e. > 2% of variants with heterozygote frequency > 0.55 (maximum expected heterozygosity is 0.5 for neutral loci). These segments generally coincided with regions of large duplications (>1 kb) that generate alignment errors and false SNP calls, therefore variants in these regions with heterozygote frequency >0.5 were removed.

RESULTS AND DISCUSSION

The pre-imputation filtering of variants in sheep Run 2 and cattle Run 6 reference sequences removed up to 25% of all variants called but this increased to 47% in Run 7, largely due to extra filters imposed. Table 1 compares the proportion of imputed variants above two Minimac R2 thresholds because the Minimac R2 statistic is a good proxy for empirical imputation accuracy (Bolormaa *et al.* 2019). The sheep imputation retained a larger number of imputed variants at Minimac R2 thresholds >0.4 and >0.8 compared to imputed cattle data. This is potentially due to the imputation target sheep having very recent relatives in the reference set compared to the cattle where relationships were more distant between the target and reference sets.

Table 1. Numbers of variants (M=Millions) imputed into sheep and cattle

No. of Variants	Sheep Run2	Cattle Run 6	Cattle Run 7
Total Imputed	40 M (77% of total)	34 M (75% of total)	32 M (53% of total)
Minimac R2 > 0.4	31 M (77% of imputed)	18 M (53% of imputed)	21 M (66% of imputed)
Minimac R2 > 0.8	22 M (55% of imputed)	14 M (41% of imputed)	19 M (59% of imputed)

Overall, only 40 to 60% of variants had a MinimacR2 >0.8. The main reason for this is due to the very high proportion of sequence variants with a MAF <0.01 (e.g. Figure 1) that are difficult to impute with accuracy above 0.8 (Pausch *et al.* 2017). Further, we hypothesise that due to purging selection, rare mutations with strong deleterious effects will tend to have arisen relatively recently, and therefore will be more difficult to impute accurately compared to rare variants that have been segregating in the population longer because they have small or neutral effects. Indeed, we found some evidence of this in both sheep (Bolormaa *et al.* 2019) and cattle where for example, missense and frameshift mutations (potentially damaging protein activity) showed a higher proportion of less accurately imputed variants compared to intergenic and intronic variants (Figure 1). In part, we may be able to improve the accuracy of imputation for rare variants by strategies such as skim whole genome sequencing (Daetwyler *et al*, these proceedings) but also by increasing the number of sequenced animals in the reference sets. An increase in the number of animals in cattle Run 7 may have helped increase the number of variants with Minimac R2 >0.8 compared to Run 6 (Table 1). However, other factors including the improved ARS-UCD1.2 reference genome map, different variant calling software and more stringent filtering of variants prior to imputation may also have contributed to the improvement and this will be further evaluated.

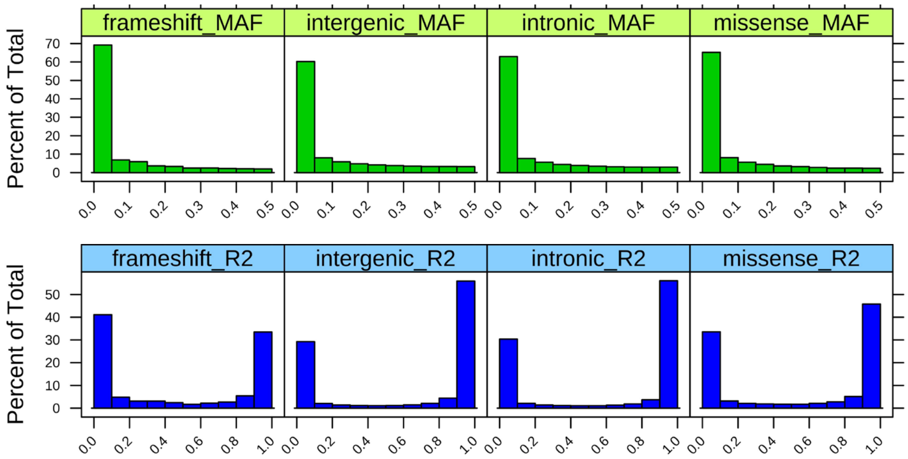

Figure 1. MAF (Minor Allele Frequency) and Minimac R2 distribution in functional categories of variants from cattle Run7. Frameshift and missense variants show the highest frequency of variants with low imputation R2

Another important factor causing low sequence imputation accuracy is an erroneous calling of SNP in the reference sequences, for example, due to alignment errors of short-read sequencing. Typically, this more frequently occurs in the many genome-wide regions of up to several Mb long that harbour large segmental repeats (each ≥ 5 kb in length) and/or large structural variants such as copy number variants (CNV) (Liu *et al.* 2010). For example, the major histocompatibility complex region has many segmental duplications and CNV (>86% synteny between cattle and sheep; Gao *et al.* 2010) and across this region the mean empirical accuracy within segments of 1 Mb length drops well below 0.8 in both sheep and cattle (Pausch *et al.* 2017; Bolormaa *et al.* 2019). In these regions, we typically observe excessive heterozygosity among reference sequence variants (i.e. heterozygosity >0.5) (Fig 2). Thus, in Run 7, prior to imputation we filtered out variants with heterozygosity >0.5 in these regions under the assumption that these are false SNP calls and may decrease the imputation accuracy of surrounding variants. As a result, on Chr X the Run 7 pre-imputation filtered variant set included only half the number of variants compared to Run 6 but the number of imputed variants in Run 7 with R2 >0.8 was almost double that of Run 6. Although stringent pre-filtering may be helpful, the low imputation accuracy of these regions (covering >3% of the genome) cannot be fully addressed with the current sequence reference sets because the short sequence reads (~150bp) cannot be accurately aligned, even though the reference genome map may be very accurate. A potential solution is to develop a reference resource where animals are sequenced using long-read technology as well as improved methods to impute large structural variants.

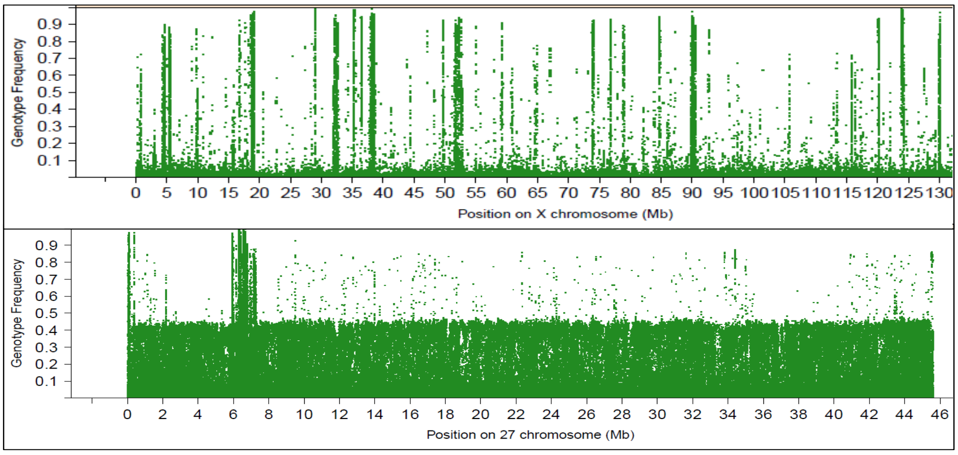

Figure 2. Frequency of heterozygous genotypes for real sequence variants on Chr X (non-pseudo autosomal region) and Chr 27. The data was derived from 2470 bulls sequenced to > 10x average read depth). Banded regions of excessive heterozygosity (>0 on Chr X and >0.5 on Chr27) coincide with large segmental repeats and copy number variants. On Chr X in addition to bands of high heterozygosity, we also observe ubiquitous random errors across the genome: i.e. these were bull X chromosome sequences that should be haploid, with "homozygous" genotypes

CONCLUSIONS

Although imputed sequence has already advanced livestock genomics research there remain considerable challenges: including rare variant imputation and limitations of short-read sequencing.

ACKNOWLEDGEMENTS

The authors acknowledge financial support from DairyBio, a joint venture between Agriculture Victoria, Dairy Australia and the Gardiner Foundation. We are grateful to the 1000 Bull Genomes Project, SheepGenomesDB, SheepCRC and DataGene for access to data.

REFERENCES

Bolormaa S., Chamberlain A.J., Khansefid M., Stothard P., Swan A.A., Mason B., Prowse-Wilkins C.P., Duijvesteijn N., Moghaddar N., van der Werf J.H., Daetwyler H.D. and MacLeod I.M. (2019) *Genet. Sel. Evol.* **51**: 1.

Bouwman A.C., Daetwyler H.D., Chamberlain A.J., et al. (2018) *Nature Genet.* **50**, 362.

Brøndum R.F., Su G., Janss L., Sahana G., Guldbrandtsen B., Boichard D. and Lund M.S. (2015) *J. Dairy Sci.* **98**: 4107.

Calus M., Bouwman A., Hickey J., Veerkamp R. & Mulder H. (2014) *Animal* **8**: 1743.

Daetwyler H.D., Brauning R., Chamberlain A.J., McWilliam S., McCulloch A., Vander Jagt C.J., Bolormaa S., Hayes B.J. and Kijas J.W. (2017) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **22**: 201.

Gao J., Liu K., Liu H., Blair H.T., Li G., Chen C., Tan P. and Ma R.Z. (2010) *BMC Gen.* **11**: 466.

Hayes B.J. and Daetwyler H.D. (2019) *Ann. Review Anim. Biosci.* **7**: 89.

Liu G.E., Hou Y., Zhu B., Cardone M.F., Jiang L., Cellamare A., Mitra A., Alexander L.J., Coutinho L.L. and Dell'Aquila M.E. (2010) *Genome Research* **20**: 693.

Moghaddar N., Khansefid M., van der Werf J.H., Bolormaa S., Duijvesteijn N., Clark S.A., Swan A.A., Daetwyler H.D. and MacLeod I.M. (2019) *Genet. Sel. Evol.* **51**: 72.

Pausch H., MacLeod I.M., Fries R., Emmerling R., Bowman P.J., Daetwyler H.D. and Goddard M.E. (2017) Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genet. Sel. Evol.* **49**: 24.

Rosen B.D., Bickhart D.M., Schnabel R.D., Koren S., Elsik C.G., Tseng E., Rowan T.N., Low W.Y., Zimin A. and Couldrey C. (2020) *GigaScience* **9**: giaa021.

Xiang R., MacLeod I.M., Daetwyler H.D., de Jong G., O'Connor E., Schrooten C., Chamberlain A.J. and Goddard M.E. (2021) *Nature Communications.* **12**: 860.