final report

Project code: j15646

Prepared by: Ben Hayes
University of Queensland

Date published: 08/10/2020

PUBLISHED BY
Meat and Livestock Australia Limited
Locked Bag 1961
NORTH SYDNEY NSW 2059

Mapping estimation and complexity of SNP coverage

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government to support the research and development detailed in this publication.

Executive summary

An analysis was conducted to identify SNPs on a number of commercially available SNP arrays widely used for genomic evaluations by the Australia beef industry in linkage disequilibrium with SNPs specifically identified by Australian Patent Application 2010202253. The accompanying excel file contains lists and ids of these SNPs.

Approach

SNP arrays currently used by the Australian beef industry include the 50k Neogen array, the 50k Illumina array, the tropBeef (35k) array, and the Illumina HD array (which includes 777k SNPs) (collectively the ***Array SNPs***).

Australian Patent Application 2010202253 specifically identifies 2510 SNPs (hereafter referred to as ***limb_a SNPs***) with relatively little overlap with the Array SNPs currently used for genomic evaluations. As a result, linkage disequilbrium (r^2) between limb_a SNPs and Array SNPs cannot be calculated from the array genotypes alone. For example, an array SNP can be in $r^2 \geq 0.65$ with a limb_a SNP, despite the limb_a SNP not being on the array.

A data set which does enable r^2 to be calculated between Array SNPs and limb_a SNPs is the 1000 bull genomes database, where all 2703 animals in the database are whole genome sequenced, and genotyped for approximately 23 million SNP. This database includes large numbers of animals from the breeds widely used in Australia, including Angus, Brahman, Hereford, Tropical Composites and many others (see Hayes and Daetwyler 2018 for full details and breed lists). This analysis has used this dataset to calculate r^2 between the limb_a SNPs and the Array SNPs.

Analysis

Mapping limb_a SNP

Limb_a SNPs are identified by Australian Patent Application 2010202253 by reference to position 300 of each of SEQ ID NOS: 19473 to 21982 disclosed in the application. The limb_a SNPs were mapped to the UMD3.1 assembly of the bovine genome (i.e. their genome coordinates were identified. A five step procedure was used as detailed below:

- Step 1. The nucleotide sequence 100 nucleotides before, and 100 nucleotides after position 300 of each of SEQ ID NOS: 19473 to 21982 (i.e. the 100 base pairs flanking each SNP) were

selected (the query sequences). The query sequences were then aligned using the BLAST algorithm Altschul et al. 1997) (to UMD3.1 ((ftp://ftp.ncbi.nlm.nih.gov/genomes/Bos_taurus/). A stringency of 4 mismatches was used (the reasons mismatches can be present between two aligned sequences can be due to errors in the sequencing of one of the aligned DNA sequences, or a natural genetic variation between the animals from which the two aligned sequences have been generated). If the algorithm finds a unique alignment between a query sequence and the bovine genome, the position of the SNP from this alignment was identified and its location recorded. The 100 nucleotide flanking the sequence for each SNP rather than the full sequence of each of SEQ ID NOS: 19473 to 21982 (approximately 600bp) because in some cases larger sequences can start entering into repeat regions, which can reduce the precision of mapping.

- Step 2. If no alignment between any one or more query sequence and the bovine genome is achieved with a stringency of 4 mismatches, the stringency was decreased incremental steps of 4 (i.e. 8, 12 etc.) and alignments were investigated. If a query sequence can be uniquely aligned with a lower stringency, this was considered the location of the sequence in the bovine genome. If the sequence aligns to multiple locations, Step 4 was performed.
- Step 3 was a quality control step. It is quite common that mutations that have initially been identified as SNPs are actually another form of a mutation known as an insertion/deletion (indel). Therefore, to make sure that the limb_a SNPs at position 300 are in fact SNPs, the sequence 100 nucleotides before position 300 and the sequence 100 nucleotide after position were aligned separately for each sequence. If the mutation at position 300 is a SNP, then the two aligned sequences will align with 1 nucleotide between them (the SNP at position 300). If, however, there is more than one position between the two independently aligned sequences, then it is likely that the mutation is not a SNP, but rather a possible indel. Indel's were retained for further analysis and effectively treated as SNP in what follows.
- Step 4. If there was no unique alignment for any of the 100 nucleotide long sequences aligned in steps 1 to 3 (i.e. they align to more than one location with the same identity), then rather than use 100 nucleotides, 300 nucleotides were used and a tolerance of 12 mismatches was set, and steps 1 to 3 were repeated. If there was a unique alignment form this step, the position of the limb_a SNP was identified and its location recorded.
- Step 5. Once the location of each of the SNPs at positon 300 of SEQ ID NOS: 19473 to 21982 were determined, the range (genomic intervals) that extends 500,000 nucleotides (500kb) was identified.

A total of **2327** limb_a SNPs mapped to both the bovine genome and were identified as either SNP or Indel in the 1000 bull genomes Run 6 (the latest run). The remaining limb-a SNPs either could not be mapped, mapped to multiple places in the bovine genome and therefore may not be true SNP, or were not segregating in the 1000 bull genomes population. This analysis is limited to the 2327 SNPs mapped to both.

Calculation of r^2 between limb_a SNP and Array SNPs

Two "populations" were extracted from the 1000 bull genomes Run 6 dataset, namely an "Aussie beef population" and an "Angus population". The Aussie beef population includes 997 animals representing breeds widely used in Australian beef production. The Angus population was an Angus only population, including 256 Angus animals.

The Arrays SNPs investigated included those on the 50k Neogen array, the 50k Illumina array, the tropBeef (35k) array, and the Illumina HD array which include 47803, 54609, 34288 and 777969 SNP respectively. These SNPs are currently widely used by the Australian beef industry for genomic evaluations of cattle. The Array SNPs and limb_a SNP genotypes were extracted from the 1000 bull genomes data set for the Aussie beef population and the Angus population separately. The level of r^2 between Array SNPs and limb_a SNPs in the Aussie beef and Angus populations was then calculated using VCF tools (Danecek *et al.* 2011). The number of Array SNPs with an r^2>=0.65 with a limb_a SNP based on these two populations was then determined.

Results.

Table 1 describes the number of ASNPs on each array that are in r^2>=0.65 with a limb_a SNP.

Table 1. Number of Array SNPs identified from the Aussie beef and Angus populations that are in r^2>=0.65 with a limb_a SNP.

	50k Neogen	50k Illumina	tropBeef (35k)	Illumina HD
Number of SNPs on chip	47803	54609	34288	777969
Number of SNPs on chip overlap with limb_a SNP	44	29	14	400
Numbers of SNPs on chip r^2>=0.65 with limb_a SNP, Aussie beef population	671	593	510	11500
Numbers of SNPs on chip r^2> =0.65 with limb_a SNP, Angus population	1574	1456	1053	23003

Lists of these SNPs are available on request.

This research was conducted using the Industry 1000 bull genome. Results may differ with different genomes.

References

Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. Annu Rev Anim Biosci. 2018 Dec 3. doi: 10.1146/annurev-animal-020518-115024.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics (Oxford, England), 27(15), 2156-8.