# final report

Project code: B.BSC.0075

Prepared by: Dr Mike Goddard
Beef CRC Limited

Date published: July 2014

ISBN: 9781740362474

PUBLISHED BY
Meat & Livestock Australia Limited
Locked Bag 991
NORTH SYDNEY NSW 2059

# Using bovine genome sequence

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

## Milestone

Sequences stored in database, phased and corrected and available for use. Investigate the potential to impute genome sequence from high density SNP genotypes. Identification of candidate mutations underlying some of the quantitative trait loci discovered in the CRC's gene discovery research eg. Net Feed Intake and horn/polled/scurs.

## Sequencing

We have sequenced 48 Angus and 32 Brahman cattle. The Angus are from the Trangie NFI selection lines. The original plan was to sequence pooled DNA from the high and low lines but we have been able to sequence the cattle individually. The depth of sequencing varies from 6 to 16 fold across animals. These animals have also had mRNA sequenced so we can examine differences in gene expression between the high and low lines. The Brahmans are from CRC I and II and have been sequenced to approximately 4 fold coverage.

## Analysis of sequence

The raw sequence reads require considerable processing before they are usable.
We (DPI Victoria and Dairy Futures CRC) have designed and developed a solution that runs sequence data through an automated software pipeline to produce genotypes at all polymorphic sites. With the primary aims of simplicity and reusing tools already widely used by the scientific community, the pipeline comprises a mix of Open Source software and scripts developed by DPI Victoria. The pipeline has three main stages: i) identification of variants in sequence (SNP and INDELs), ii) filtering of variants to reduce erroneous variant calls, and iii) genotype correction using haplotyping software. Currently sequence variants are identified using the samtools mpileup software, which simultaneously analyses the sequence reads of a set of animals. These variants are then filtered using samtools-varfilter and vcftools to remove calls which are based on too few observations in the sequence, are of low quality, or are likely due to systemic sequencing errors. The remaining variants are then analysed with a haplotype program (in our case Beagle), which corrects genotypes based on the probability of the genotype and on the haplotypes of other animals in the sample. A number of automated quality control steps have been implemented which include checking opposite homozygotes in parent offspring pairs and checks against 800k SNP chip data. This enables both checking of input data and how well the pipeline is performing. To meet software runtime performance requirements from the terabytes of data involved, the pipeline is parallelised to make use of DPI's computational cluster and proven job queueing software. This process results in sequences that are corrected and phased.

## Imputation of genome sequence from SNP genotypes

We (Ben Hayes et al) have carried out a proof of principle using Holstein genome sequence and SNP genotypes. Using 12 sequenced bulls we imputed full sequence on cattle with SNP genotypes in the region of the KIT gene. The accuracy of imputation varied across sites from random to 99% accurate. The low accuracy occurs when a haplotype in the SNP data has not been observed in the animals that have been sequenced. This problem can be overcome by sequencing more animals. However, to impute sequence into rare haplotypes will require a large database of sequence and this is why we have initiated the 1000 bull genomes project. We had previously discovered that a quantitative trait locus (QTL) affecting the proportion of white on the coat of Holsteins maps to the KIT gene. We used the imputed sequence data to search for mutations that are associated with white spotting. We found 3 SNPs that are associated with white spotting suggesting that there are multiple alleles at this gene.

We have not yet carried out a similar analysis with the beef cattle data because we are waiting for the first analysis of the 1000 genomes database scheduled for February.

## Use of the sequence to find causative mutations

We have mapped QTL for carcase and meat quality, net feed intake and fertility using genotypes on 700,000 SNPs from 10,000 cattle. Several genome regions are highly significant (p<0.000001). We will examine the full sequence in these regions and impute sequence on animals with SNP genotypes in March after the 1000 genome analysis.

## Future use of sequence data

Beef CRC has entered into two collaborations to increase the usefulness of our sequence data. Firstly, we are part of a Genome Canada project that will sequence many cattle and share the sequence. Secondly, we are collaborating with DPI's "1000 bull genome "project. DPI and the Dairy Futures CRC are hosting this project which aims to collect 1000 bull genome sequences from institutions around the world. These sequences will be corrected and phased using the software described above and will then be available for imputation of sequence from SNP genotypes. The accuracy of imputation depends on the number of sequences available and so collaboration benefits all partners. We expect the Genome Canada sequences will be added to the database when they are available. We will run the first analysis of the database in February 2012 when we expect to have >200 cattle genomes in the database. By 2013 we hope to reach the target of 1000 genomes.

## Implications for beef cattle research

Genomic selection based on 50,000 SNPs is very successful in Holstein cattle provided a large database of cattle with genotypes and phenotypes can be used to derive the prediction equations. For instance, the USDA uses 16,000 Holstein bulls all of whom have been genotyped and have been progeny tested. To achieve the same accuracy of genomic selection in beef cattle based on individual cattle measurements we would need approximately 48,000 steers within each breed. This is not possible in the foreseeable future. The alternative is to utilise the data from all breeds to derive a prediction equation that applies to all breeds and this is the goal of the CRC.

However, to achieve this we need to find mutations that cause variation in economic traits or markers so close to them that they are consistently associated with the causal mutation in all breeds (ie in high linkage disequilibrium or LD). The 800k SNP chip is useful in this regard. However, causal mutations which are comparatively rare will not be in high LD with SNPs on the commercial SNP chips all of which have been chosen to be common. The ideal solution is to replace SNP genotype data with genome sequence because the causal mutations will be present in the full sequence and therefore we don't have to rely on LD between causal mutations and SNPs.

The use of genome sequence was not possible until recently. In the last decade the cost of genome sequencing has dropped 1,000,000 fold. However, it is still not possible to use sequencing directly instead of SNP genotyping due to cost and time. Therefore, we have developed a strategy to impute full genome sequence on animals with SNP genotypes using a reference set of animals that have full sequence. We have demonstrated this procedure in a part of the genome such as the kit locus described above.

The strategy requires a large number of animals with full genome sequence in the reference database but this set of animals could be used by all researchers around the world and consequently we have proposed and implemented the 1000 cattle genomes project. The contributions of Angus and Brahman that this project has made to the database are important because at present they are the only Angus and Brahman cattle included.

Two additional problems must be overcome to utilise this strategy of imputing genome sequence. The traditional standard to determine the sequence of an animal has been 30 fold coverage of the genome because with low coverage the genotype called at a given position in the sequence for a particular animal may be incorrect. This is still too costly. Therefore, we have used the haplotype of sequence variants in a small region of the genome to help determine the genotype of an animal at each site within that haplotype. That has made it possible to infer the sequence on animals that have only been sequenced at an average coverage of 4 times.

The second problem with the use of genome sequence is simply the size of the data. Each animal that is sequenced generates 0.5 terabytes of data. Few research groups have the computing resources or scientists trained to handle such data. DPI Victoria has invested in both. Via the 1000 bull genomes project it is now possible for other researchers to submit genome sequence and have it processed to correct errors and call genotypes and haplotypes. It is also possible to submit SNP genotypes to the database and have full sequence imputed.

At first the labour requirement to do this was very high – it took weeks for each animal that was sequenced. However, the pipeline that has been built has greatly speeded up this process so that it is now semi-automatic but still requires some expert intervention. This greatly reduces the bioinformatic cost of utilising sequence data in cattle research.

## Implications for genetic improvement of beef cattle

Genomic selection has already been implemented world wide in dairy cattle and is doubling the rate of genetic improvement. As argued above, the exact strategy used for Holstein is less applicable to beef cattle. For major mutants, such as those causing Mendelian genetic abnormalities, finding the causal mutation and testing for that mutant allele is the normal approach. There are several reasons to think that this approach will be beneficial for quantitative traits as well.

Mutants causing genetic abnormalities are usually rare. It is likely that mutations with more modest effects are also rare in the absence of selection for them. For instance, there are several known mutations at the myostatin gene that cause double muscling but there is also the F94L mutation in Limousin which has a more modest effect. Mutations such as F94L might be rare and undetected by a BLUP analysis using 50k SNP data but detected using genome sequence data.

Even when causal variants are not rare, simulation shows that including the causal variant in the data increases the accuracy of genomic selection (Meuwissen and Goddard 2010, *Genetics* **185**:623-631).

Thus we expect that use of imputed genome sequence data will make possible our goal of a prediction equation based on data from all breeds that is accurate in all breeds.

In addition, where it is possible to find the causal variant, or at least identify the gene involved, this will increase our understanding of the biology underlying genetic variation and this will have additional benefits such as predicting GxE interactions.