APPLICATION OF WHOLE GENOME SEQUENCE TECHNOLOGY TO DAIRY CATTLE BREEDING BY LIC.

M. Keehan, A. Scott, T. Lopdell, T. Johnson and R. Spelman

LIC, Private Bag 3016, Hamilton 3240, New Zealand

SUMMARY

Whole genome sequence (WGS) technology has become affordable to animal breeding companies. This paper will describe how LIC is applying WGS to dairy cattle breeding, show results LIC has currently obtained from WGS and the expected benefits for dairy farmers. The benefits to farmers are improved reliability of genomic evaluation and the ability to detect and control low frequency recessive variations of large effect within the NZ dairy population.

INTRODUCTION

LIC is an artificial breeding and herd recording dairy cooperative owned by farmer customers that has approximately 80% of the NZ dairy market. LIC has been marketing bulls selected by genomic evaluation as well as the traditional method of evaluation based on daughter performance. Teams of genomically selected bulls have lower reliability than daughter proven bulls but they can improve the overall rate of genetic gain due to the lower generation interval. A genomically selected bull can be used at one or two years of age whereas a daughter proven bull is widely used at five years of age. This increase in the overall rate of genetic gain has potentially huge benefits for New Zealand. Despite these benefits there is huge customer demand to increase the accuracy and reduce the bias of genomic evaluations.

LIC has an extensive resource of genotypes from commercial Single Nucleotide Polymorphism (SNP) chips for gene discovery and genomic evaluation. They include 60,000 animals genotyped on the 50 thousand (K) SNP Illumina Bovine SNP50 (Illumina Inc., San Diego CA), 3276 animals genotyped on the 700K SNP Illumina Bovine HD and 12,000 on the “GGP” 9K SNP Geneseek Genomic Profiler (Neogen Corp., Lincoln, NE). Recently it has become cost effective to augment a standard SNP chip with additional custom SNP content.

The cost of whole genome sequence technology has declined from billions of dollars for the first human genome sequence to less than US$5000 per sequence. This tremendous reduction has been achieved from major improvements in all areas of genome technology. The decrease has enabled the 1000 Genomes Project Consortium (2010) to pioneer widespread low coverage sequencing to discover low to rare frequency variants in the human population. The 1000 Genomes project improved many bioinformatics tools e.g. alignment algorithms (Li and Durbin 2010), Samtools (Li *et al.* 2009) and variant detection and calling - GATK (Depristo *et al.* 2011). Phasing of genotypes from low coverage sequence data to form a reference panel which can then be used to impute sequence data from a low cost SNP chip is another important advance.

Whole genome sequence (WGS) has advantages over commercial SNP chips. Most variants in the population are found including structural changes, insertion, deletions and variants excluded by SNP chip chemistry. In some circumstances the maternal or paternal phase of the SNPs can be determined from WGS. Disadvantages of WGS include data size and comparatively low accuracy at low sequence coverage compared to SNP chip genotypes. Levels of sequence coverage can also vary throughout the genome.

LIC has been funded to embark on a seven year research project to apply WGS technology to dairy cattle breeding within NZ. This paper reports how LIC has been utilising the WGS data for both genomic selection and gene discovery and gives preliminary and ongoing results from analysis of the sequence data and the development of a bovine mammary transcriptome resource.

MATERIALS AND METHODS

Sampling. Tissue samples have been taken from 502 animals. Phase 1 consisted of 25 animals sequenced in 2011 and Phase 2 includes an additional 477 sequenced in late December 2012. Samples came from the LIC quantitative trait loci gene discovery herd, LIC bull semen stocks and the Vialactia phenotypic outlier discovery program. Key ancestral bulls from the national pedigree were chosen by the program ExomePicks (Abecasis 2010).

Library construction. DNA was extracted and sent to Illumina FastTrack for library construction. Most libraries generated were 100 base pair paired end reads. 2K, 5K and 10K insert size libraries were constructed from one high coverage animal intended for deNovo assembly.

Sequencing. All libraries were sequenced on Illumina HiSeq. Twenty three phase1 animals sequenced in 2011 ran on the V2 chemistry and were sequenced to an average 28X coverage. The 477 remaining animals ran on the newer IlluminaV3 chemistry.

Mapping. Reads were aligned using the RTG version 2.7.2 (Real Time Genomics, Inc. San Bruno, CA) against version 3.1 of the University of Maryland Bovine genome reference assembly. (Zimin *et al* 2009) . Median mapping coverage (X) is 6X, mean is 10X and maximum is 138X.

Genotype Calling. Genotypes were called using RTG, Samtools and GATK pipelines. Phase 1 genotypes were called using the consensus from the three pipelines. Phase 2 genotypes were called using RTG v 2.7.2 in pedigree aware mode and samtools in multiple sample mode with default parameters. Phase 2 consensus genotypes have not been called at time of writing.

Imputation. An early trial assessment of imputing sequence has been done using Beagle version 3.3.2 (Browning 2006). The first 4 Mb of chromosome 3 from the RTG pipeline was selected as it had completed genotype calling first. SNP Variants to impute were selected by requiring a variant call format QUAL field > 30. The sequence genotypes were sub sampled to either the 50K or GGP set of markers. The subsampled genotypes were then imputed back to 50K, HD and then Sequence. The 50K and HD reference were the LIC standard production reference populations of 15,000 animals genotyped with 50K and 3000 with HD SNP chip.

Deletorious recessive detection. The program SnpSift from snpEff (Cingolani et al 2012) was used to identify variants that lack observed homozygotes of one allele. These variants were further filtered by snpEff and manual curation. Fourty nine candidates were then sent for validation genotyping in a sample of 1350 animals.

Genomic Selection. The Phase1 dataset was used to generate a list of SNPs within 32 important dairy production genes chosen from the scientific literature. This list of 500 SNPs was added as LIC custom content to a Neogen GGP panel. 13,650 animals from the LIC Sire Proving Scheme will be genotyped on this custom list. Results have not been analysed at the time of writing.

RNA-Seq. Bovine mammary samples were extracted from 29 cows. RNA was converted to DNA at the University of Auckland. NZ Genomics performed sequencing on Illumina HighSeq. Results were analysed using a standard Tophat and Cufflinks pipeline. (Trapnell *et al.* 2009)

RESULTS AND DISCUSSION

SNP Discovery. 29,362,664 SNP were called from the Phase 2 RTG call set.

Genotype calling. 62 of the 502 sequenced animals have independent HD genotypes and 393 have independent 50K genotypes. Concordance has been computed between the sequence and HD genotypes and varied depending on the average level of mapped coverage per animal. For sequence coverage of 4X to 6X concordance was around 97%, at 10X concordance rose to 99% and at >20X concordance was greater than 99.5%

Imputation. Imputing a 4 megabase test region of the genome from GGP and 50K to sequence gave a genotypic concordance of 0.93 and 0.95. The sequence reference was then filtered to use only SNPs with a reference phasing beagle $R^2 > 0.995$. This filtering removed 75% of the SNPs and gave genotypic concordances of 0.94 and 0.95. The "round trip" validation method (subsampling and using self as a reference) represents a perfect haplotype distribution match between the reference and the imputing population and gives an upper bound imputation accuracy Johnson *et al.* (2011) reported concordances of 0.98 to 0.99 for imputing from 50K to HD. The initial low imputation accuracy of sequence can be attributed to errors in heterozygotes from low coverage genotypes, and enrichment of sequence errors within low allele frequency SNPs. Future work will improve imputation accuracy by better SNP and genotype filtering strategies, using additional phase information from reads and pedigree and genotype likelihood recalibration techniques. Currently the whole genome call set is 120 GB for the 500 animals. If the 50,000 animals genotyped on the 50K SNP chip were to be imputed to sequence it would generate a 12 TB file. The exploitation of a 12 TB file is computationally and statistically challenging.

Deletorious recessive detection. Analysis of the validation dataset has not yet been completed. It was observed that one of these predicted recessives was also in the centre of an association peak for a currently unpublished recessive mutation and explained the phenotype perfectly. Hence it is likely this will be a powerful technique to detect recessive acting genes of large deleterious effect and low to moderate frequency in the population. This approach is similar to the method used by Van Raden *et al.* (2011) to discover three Holstein fertility haplotypes in the North American population. Farmers should expect further improvements in fertility through the reduction in frequency of deleterious alleles.

Genomic Selection. Genomic selection theory is based on using a panel of anonymous evenly spaced markers that are individually in linkage disequilibrium with causative variants. Whole genome sequence technology enables the discovery of almost all variants within the population of interest. Thus in theory it should be possible to identify all causative variants or a sufficiently large subset of markers with the greatest LD to the causative variant Unpublished LIC internal experiments with 2 major milk protein markers as fixed effects have shown improvements to genomic evaluation which suggests the benefit from using causal markers.

RNA-seq. The first application of the RNA-seq data has been used to correct annotation within a genomic interval that is being investigated as a causative gene. An additional exon was discovered in that region and the presence of a 3 base pair in-del confirmed.

CONCLUSION

The WGS data is being utilised in many ways. The first use of the sequence data has been as a NZ population specific SNP discovery platform. These SNPs will be used in genomic evaluation and to search for additional recessive alleles reducing fertility. RNA-seq will be used indirectly to help explain and validate the actions of causative variants. Imputation of WGS will be restricted to small regions of the genome or subsets of animals until imputation accuracy has been improved.

ACKNOWLEDGEMENTS

This research was funded by the New Zealand Ministry for Primary Industries through the Primary Growth Partnership and LIC.

REFERENCES

Abecasis G. (2010) http://genome.sph.umich.edu/wiki/ExomePicks

Browning S. (2006) *Am J Hum Genet* **78**:903.

Cingolani P., Platts A., Wang le L., Coon M., Nguyen T., Wang L., Land S.J., Lu X., Ruden D.M.

(2012) *Fly* **6:** 80.

DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernytsky A.M., Sivachenko A.Y., Cibulskis K., Gabriel S.B., Altshuler D. and., Daly M.J. (2011*) Nat Genet* **43**: 491.

Johnson D., Spelman R., Hayr M. and Keehan M. (2011) *Proc. Assoc. Advmt. Anim. Breed. Genet*. **19**:379.

Li H., Handsaker B., Wysoker A., Fennel T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. (2009) *Bioinformatics* **25**: 2078.

Li H. and Durbin R. (2010) *Bioinformatics* **26**: 589.

The 1000 Genomes Project Consortium (2010) *Nature* **467**:1061.

Trapnell C., Pachter L., Salzberg S.L. (2009) *Bioinformatics* doi:**10**:1093/bioinformatics/btp120

VanRaden P., Olson K., Null D. and Hutchison J. (2011) *J Dairy Sci* **94**: 6153.

Zimin A., Delcher A., Florea L., Kelley D., Schatz M., Puiu D., Hanrahan .F, Pertea G., Van Tassell C., Sonstegard T., Marcais G., Roberts M., Subramanian P., Yorke J. and Salzberg S. (2009) *Genome Biology* **10**:R42