Using RNAseq data to improve genomic selection in dairy cattle

T. Lopdell[1,2] K. Tiplady[1] & M. Littlejohn[1]

[1] R&D, Livestock Improvement Corporation, Ruakura Rd, Newstead, Hamilton, New Zealand
[2] School of Biological Sciences, University of Auckland, Symonds St, Auckland, New Zealand

Summary

Genomic selection is playing an increasingly important role in animal breeding worldwide. For reasons of cost and computational feasibility, it is useful to select variant sets of maximum predictive ability, while minimising the number of variants required. RNA sequencing has the potential to aid variant selection in two ways: first, by enabling the discovery of variants of strong biological relevance (by virtue of their expression in tissues of interest); and second, by empowering the discovery of expression QTL, enabling enrichment of variant data with loci of demonstrable, modulatory effect. In this study, RNAseq was performed on lactating mammary gland from 373 New Zealand dairy cattle, followed by variant calling and eQTL discovery. Significant eQTL were identified at 3,738 genes, yielding 3,695 distinct tag variants, which were subsequently tested for their ability to predict milk volume, fat and protein phenotypes in a genomic selection context. We show that variants selected in this manner show good predictive abilities for these phenotypes, and that RNAseq is a potentially useful approach for enhancing variant selection.

Keywords: RNAseq, genomic selection, SNPs, SNP calling

Introduction

High-throughput sequencing of RNA (RNAseq) is the current gold-standard approach for evaluating gene expression levels, and can be used to discover expression quantitative trait loci (eQTL). Unlike other technologies such as RT-PCR or microarrays, RNAseq also provides additional data which is potentially useful for genomic evaluation: variant detection and genotyping. One approach to improving the accuracy of genomic selection is to find a set of variants with improved predictive power. The aim of this work was to use RNAseq from lactating mammary glands to facilitate the selection of a set of predictive SNPs for dairy cattle, in two ways: first, by enabling the discovery and genotyping of dense clusters of variants within expressed regions of the genome; and second, by detecting eQTL to identify tag variants of modulatory loci that may also impact complex phenotypes.

Materials and Methods

RNA sequencing and bioinformatics

RNA sequencing was performed on lactating mammary gland biopsies from 373 mixed-breed (Holstein-Friesian, Jersey, and crosses) mixed-age New Zealand dairy cows (Littlejohn *et al.*, 2016). Samples were sequenced using Illumina HiSeq 2000 instruments, producing 100 bp paired-end reads. Reads were mapped to the bovine UMD 3.1 reference genome using Tophat2 (version 2.0.12; Kim *et al.*, 2013). Gene expression levels were counted for

reads mapping to exons in the Ensembl gene annotation (release 81). Reads counts were normalised using the variance stabilising transformation (VST) in DESeq (version 1.28.0; Anders & Huber, 2010) to produce phenotypes for eQTL mapping. SNP calling was performed using GATK HaplotypeCaller (version 3.3.0; McKenna *et al.*, 2010) and Samtools/BCFTools (version 1.1; Li, 2011). Variant calls from each method were phased using Beagle (version 4.0; Browning & Browning, 2007) with ten burn-in iterations and ten phasing iterations, then filtered to exclude those with: total read depth <8, allelic R2 <0.95, call rate <0.9, and alternative allele frequency <0.025. The intersection of the two filtered SNP sets was phased again, using Beagle as described above, to generate the final RNAseq variant set. All animals were also genotyped on the Illumina BovineHD SNP-chip.

eQTL mapping and tag SNPs

The RNA variant set was subsequently merged with the BovineHD genotypes to provide a scaffold of variants with an approximately uniform spread across the genome. Gene expressions were analysed using Matrix-eQTL (version 2.1.1; Shabalin, 2012) to identify genes with significant *cis*-eQTL in the mammary gland. To account for population stratification, covariates were calculated from the HD genotypes using the "mds-plot" method in Plink (version 1.90b3i, Chang *et al.,* 2015). The first ten components were fitted as covariates in Matrix-eQTL. Genes were considered to have a significant *cis*-eQTL when at least one SNP within 500kb had a p-value smaller than 1×10^{-5}. From each of these genes (n=3,738), the most highly associated SNP was chosen as the tag variant. In the case of equally significant variants, the variant closest to the start of the gene was chosen, based on the Ensembl annotation (release 81). Results were then aggregated across genes, yielding 3,695 distinct tag variants (the discrepancy in number due to a subset of neighbouring genes sharing variants).

Predictive ability for genomic selection

The 3,695 tag variants were imputed into a training population of 4,982 bulls that were born prior to 2009, and a test population of 331 bulls born from 2009 to 2011. Yield deviation phenotypes were created for milk yield, fat yield, and protein yield, using daughter herd test results, with adjustments for contemporary group, age at calving, and month of calving relative to the herd start of calving. Training and testing were performed using a weighted BayesB model in GenSel (version 4.53R; Habier *et al.*, 2011) with pi=0.95, 10k burn-in iterations, and 40k sampling iterations. The test set was evaluated by calculating the correlations between the predicted BVs in the test set with the daughter-proven BVs for the same bulls.

Results

Variant calling and concordance with HD genotypes

The final RNAseq variant set contained 477,531 variants. Adding a filtered subset of the BovineHD variants (675,321) yielded a final HD+RNA variant set consisting of 1,093,581 variants. For variants in common between the HD and RNA sets, overall genotype concordance was 0.988, with a non-reference sensitivity of 0.993 and a non-reference discrepancy of 0.015.

Numerous significant eQTL detected

A large number of genes (3,738) exhibited *cis*-eQTL where the top associated variant had an association stronger than $1x10^{-5}$. These include several genes that are known to be associated with milk production and composition phenotypes, including GPAT4 (Littlejohn *et al.*, 2014; $P=7.19\times10^{-21}$), DGAT1 (Grisart *et al.*, 2002; $P=3.77\times10^{-53}$), MGST1 (Littlejohn *et al.*, 2016; $P=1.90\times10^{-54}$) and PLAG1 (Fink *et al.*, 2017; $P=1.47\times10^{-14}$).

Variant selection for GS

Correlations between predicted and daughter-proven BVs were 0.389 for fat yield, 0.510 for protein yield, and 0.690 for milk yield for the 3,695 tag variants. To test whether these variants improved prediction accuracy, random samples (100 per phenotype) were chosen from a 34k filtered subset of Illumina Bovine50k SNP chip genotypes and prediction accuracies calculated. These are plotted below in Figure 1. Predictions from tag variants for protein and milk yield exceeded those from any of the random samples, while predictions for fat yield were at the 98th percentile. For comparison purposes, the prediction accuracies for the full 34k variant set are also plotted, and exceed all predictions based on smaller variant sets.

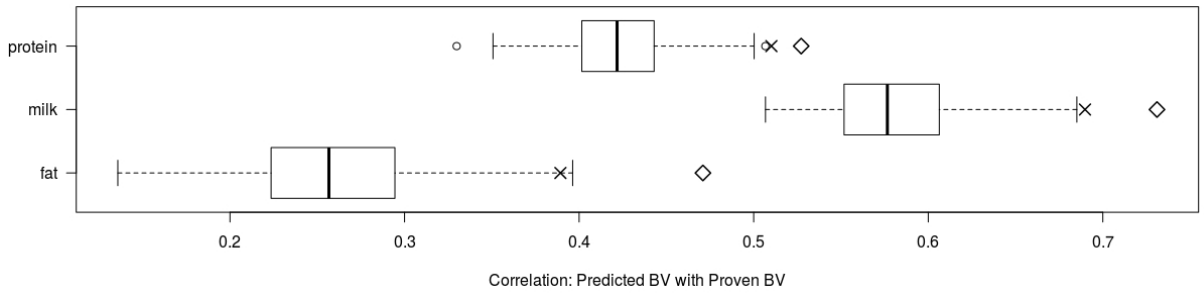

Figure 1. Ranges of correlations for predicted BVs for milk, fat, and protein yield using 3,695 random SNPs from the Illumina 50k bovine SNP chip, compared to daughter-proven BVs. Crosses mark the prediction accuracy for the eQTL tag variants for each phenotype. Diamonds mark the prediction accuracies for the full 34k variant set.

The most highly predictive region for all three phenotypes was the 1Mbp region surrounding the DGAT1 locus, where three variants explained 22–46% of the SNP variance in the tag-variant set. Other highly predictive regions included the MGST1 (Littlejohn *et al.*, 2016) and GPAT4 (Littlejohn *et al.*, 2014) loci.

Discussion

The variant set used for GS has a bearing on both the predictive ability and computational efficiency of the process. Here, we have used RNAseq data from lactating bovine mammary glands to generate a set of cis-eQTL tag variants. These variants are necessarily located in or near genes that are expressed in the mammary gland, and, by virtue of tagging an eQTL, have proven biological function. Here, we have shown that these variants also have above-average predictive ability for lactation phenotypes. Despite accuracies surpassing those of random SNPs, the predictions based on eQTL tag variants did not exceed $R^2>0.5$ for any phenotype. One reason is that not all QTL are caused by underlying expression effects. For

example, the Y581S mutation in the ABCG2 gene (Cohen-Zinder *et al.*, 2005) is known to be associated with milk yield, but was not included in this study, as the ABCG2 gene lacks an eQTL. Other loci may have eQTL in different tissues or at different stages of development. It is therefore likely that prediction would be improved by integrating additional variants from multiple sources, an idea supported by the higher prediction accuracies produced by the 34k variant set which, despite containing no causative variants, has sufficient density for biologically active loci to be picked up by linkage disequilibrium. In conclusion, we have shown that predictions for genomic selection can potentially be improved by incorporating RNAseq data, by enabling variant calling in biologically relevant parts of the genome, and by prioritising variants based their ability to tag eQTL as a proxy for biological impact.

List of References

Anders, A., W. Huber, 2010. Differential expression analysis for sequence count data. Genome Biology 11:R106.

Browning, S.R., B.L. Browning, 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097.

Chang, C. C., C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7.

Cohen-Zinder, M., E. Seroussi, D. M. Larkin, J. J. Loor, A. Everts-van der Wind, *et al.*, 2005.
Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Research 15:936–944.

Fink, T, K. Tiplady, T. Lopdell, T. Johnson, R. G. Snell, *et al.*, 2017. Functional confirmation of PLAG1 as the candidate causative gene underlying major pleiotropic effects on body weight and milk characteristics. Scientific Reports 7:44793.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, *et al.*, 2002. Positional Candidate Cloning of a QTL in Dairy Cattle: Identification of a Missense Mutation in the Bovine DGAT1 Gene with Major Effect on Milk Yield and Composition. Genome Research 12:222–231.

Habier, D., R. L. Fernando, K. Kizilkaya, D. J. Garrick, 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Kim, D., G. Pertea, C. Trapnell, H. Pimental, R. Kelley, S.L. Salzberg, 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 14:R36.

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987–2993.

Littlejohn, M. D., K. Tiplady, T. Lopdell, T. A. Law, A. Scott, *et al.*, 2014. Expression Variants of the Lipogenic AGPAT6 Gene Affect Diverse Milk Composition Phenotypes in *Bos taurus*. PLoS ONE 9(1):e85757.

Littlejohn, M. D., K. Tiplady, T. A. Fink, K. Lehnert, T. Lopdell, *et al.*, 2016. Sequence-based Association Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition. Scientific Reports 6:25376.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulski, *et al.*, 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing

data. Genome Research 20:1297–1303.
Shabalin, A. A, 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28 (10):1353–1358.