A Bayesian Analysis to Exploit Imputed Sequence Variants for QTL discovery.

I.M. MacLeod[1*], B.J. Hayes[2,3], C.J. Vander Jagt[3], K.E. Kemper[1], M. Haile-Mariam[3], P.J. Bowman[3], Chris Schrooten[4] and M.E. Goddard[1,3].

[1]Melbourne School of Land & Environment, University of Melbourne, [2]Biosciences Research Centre, La Trobe University, [3] Department of Environment & Primary Industries, Melbourne, Victoria, Australia. [4]CRV, 6800 AL, Arnhem, Netherlands. [*]Dairy Futures Cooperative Research Centre, La Trobe University, Bundoora VIC 3083, Australia

ABSTRACT: Bayesian genomic prediction methods, commonly used for genomic selection in livestock, are potentially a powerful tool for QTL discovery. However these methods become more computationally challenging as we move from HD SNP to sequence variants. We discuss results from a modified BayesR analysis in which 800K genotypes and subsets of imputed sequence variants were allocated to specific classes based on biological information prior to starting the analysis. The analysis determines if there is enrichment for QTL effects by allowing the distribution of SNP effects to vary between classes. We analysed milk traits from a mixed group of Holstein and Jersey bulls. Using examples of mutations in genes previously associated with milk traits, we demonstrate that this modified Bayesian analysis may provide a powerful approach for short-listing potential causal variants.

Keywords: BayesRC; dairy cattle; causal mutations

Introduction

Bayesian methods such as BayesR (Erbe et al. 2012) are often used as a tool for genomic selection of livestock based on SNP genotypes. However, these methods can also be used as a tool for QTL discovery. BayesR for example, simultaneously predicts genome-wide SNP effects and calculates a posterior probability for each SNP indicating the likelihood that it has a real effect on the trait. Potential advantages of Bayesian approaches over other QTL mapping methods, such as single marker regression commonly used in GWAS are: they fit all SNP simultaneously in a single model, they allow for a flexible distribution of SNP effects so that many SNP have a very small or zero effect while others have a large effect. A new feature of these Bayesian models is introduced in this paper, that is, the possibility of utilising biological information in the model priors.

In dairy cattle breeds such as Jersey and Holstein, the long range linkage disequilibrium (LD) is high which means that quite a number of SNP may show a strong association with a QTL. Although Bayesian methods are better able to deal with this than GWAS, by fitting all SNP simultaneously, this issue will increase as we move from HD SNP to sequence data.

In this study we introduce a modified BayesR method, BayesRC, which we developed to take advantage of biological knowledge that is available a priori on sequence variants. We compare BayesR and BayesRC methods for QTL discovery in dairy cattle data, using both HD 800K SNP genotypes as well as subsets of imputed sequence data. We combined data from Jersey and Holstein cattle in an effort to reduce the number of SNP tagging each QTL.

Materials and Methods

Genotype Data. We obtained either real or imputed 800K HD SNP data (Illumina) for 6920 Holstein bulls and 1108 Jersey bulls. For this study we identified three subsets of genome-wide sequence variants. The first was a set of variants in regions predicted to have a potentially regulatory role (REG), such miRNA variants and regions just up or down-stream of genes. The second was all non-synonymous coding (NSCoding) variants (i.e. DNA base pair modifications that change an amino acid in the protein code) and the third was all 800K SNP not included in REG or NScoding sets. The first two subsets of sequence variants were imputed for all 8028 bulls using sequences from the 1000 Bull Genomes Consortium (http://www.1000bullgenomes.com/). Here we refer to the three sets of variants as REG, NSCoding and 800K, while all three sets combined are referred to as SEQ. In the SEQ data if a pair of variants were in perfect LD (r^2>0.99) then one of the pair was removed, and all monomorphic variants were discarded. In total 994,019 variants remained, of which 45,026 were NSCoding, 587,734 REG and 387,170 800K.

Milk trait records. We calculated de-regressed proofs for all bulls in the study from their international MACE dairy bull breeding values. Traits included milk, fat and protein yield (Milk, Fat & Protein).

BayesR and BayesRC Statistical methods. BayesR methodology was implemented here as detailed in (Erbe et al. 2012) except that we included a correction for breed as a fixed effect in the model. Bayes R assumes that the effects of SNP are drawn from a mixture of normal distributions each with a mean of zero and variance of: (i) zero, (ii) $0.0001\sigma_g$, (iii) $0.001\sigma_g$ or (iv) $0.01\sigma_g$, where σ_g is the additive genetic variance of the trait. The analysis estimates the proportion of SNP that fall into each of these classes. The methodology for BayesRC is the same as BayesR except that, a priori, each SNP (variant) is identified as belonging to a specific "class" (2 or more) and the proportion of SNP that belong to each of the four normal distributions can vary between classes. Therefore the class to which a SNP belongs can affect the posterior probability that it has a non-zero effect. Consequently, SNP that belong to a class that on average has a large effect on a trait, are more likely to be included in the model than SNP that belong to a class that rarely has an effect on the trait.

Class allocation was based on the predicted genomic properties of sequence data (REG or NSCoding)

as well as on lactation biology. For the latter, we defined a set of 792 genes which were differentially expressed in the mammary gland in response to treatments that altered milk yield, from an independent study (Vander Jagt 2012). All variants in our genotypes that were in or within 50Kb of these genes were defined as "Lact".

We carried out four different analyses:

1. BayesR 800K: BayesR using only 800K SNP data.
2. BayesR_SEQ: BayesR using all SEQ variants.
3. BayesRC_SEQ: BayesRC using SEQ variants: class I = NSCoding, class II = REG, class III = 800K i.e. all other SNP in SEQ.
4. BayesRC_Lact: BayesRC using SEQ variants: class I = all NSCoding variants that were located in the Lact genes, classII = all REG and 800K variants in Lact genes, class III = all other SEQ variants outside the Lact set.

Results and Discussion

SEQ variants. We calculated the frequency of variants in each of the 3 sets, and found that more that 50% of the NSCoding variants had a minor allele frequency (MAF) of ≤ 0.1, while the majority of 800K SNP had MAF >0.1 (Fig. 1). The likely explanation for this large difference is that 800K SNP are chosen to have MAF>0.1 while NSCoding variants often cause strongly deleterious effects on fitness traits and therefore are selected to low frequency. This suggests that LD will often be low between these potential causal mutations and the 800K SNP because the latter are specifically chosen for their high MAF. Assuming a QTL with MAF = 0.05, then to achieve a minimum r^2 of 0.5 between a SNP this QTL, the SNP allele frequency must be ≤ 0.1 (Wray 2005).

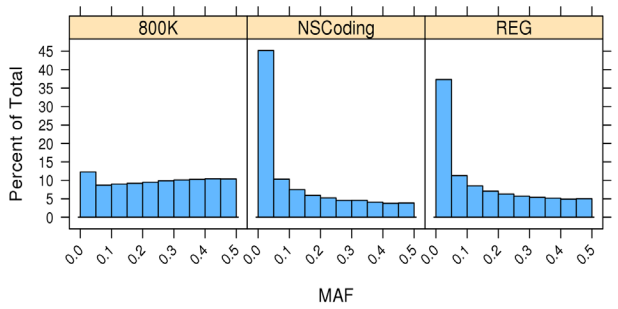

Figure 1. MAF distribution among the different subsets of variants used in the analysis.

In this case it is expected that it will be difficult to detected the effect of low frequency causal mutations on a polygenic trait unless the actual mutation is present in the data. By including all imputed NSCoding variants in the analyses we hoped this set would be enriched for causal variants affecting milk traits.

The REG set included all variants in 2Kb regions up and down-stream of known genes so, although there would undoubtedly be some causal variants present, a large proportion of our REG variants will have no effect on any trait. Therefore it follows that their MAF distribution is less extreme than the NSCoding variants.

QTL Discovery. Figure 2 shows the results for QTL analysis in the well-known DGAT1 region of chromosome 14. We compare the results for Milk with 4 different analyses: 800K_BayesR, Lact_bayesR, Seq_BayesR and Seq_BayesRC. The previously identified DGAT1 NSCoding causal mutation was not present among the 800K SNP but there was a single SNP in strong LD with the DGAT1 mutation and which was included in the model every iteration (that is, it had a posterior probability of 1 of affecting milk yield). There is a scatter of other SNP close by that also show moderate posterior probabilities indicating that the 800K SNP did not capture all of the variance. DGAT1 was included in the Lact gene set and therefore the DGAT1 causal mutation was assigned to Class I variants in the Lact_BayesRC analysis. This analysis clearly "discovers" the causal mutation with a posterior probability of 1. The BayesR_Seq analysis does not use any prior biological information and was not able to distinguish the causal DGAT1 mutation from among a number of other SNP around the DGAT1 gene that were in moderate to high LD with the causal mutation.

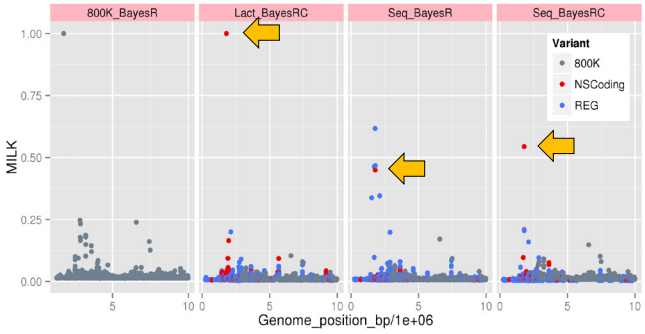

Figure 2. Posterior probabilities for a range of BayesR and BayesRC analyses of Milk Yield on a region of chromosome 14 highlighting the known DGAT mutation (arrow).

The fourth analysis shown in Fig. 2 is Seq_BayesRC for which the DGAT1 mutation was also in Class I. However in this case the posterior probability for the mutation is lower than the Lact_BayesRC. The likely explanation is that Class I variants in the Lact_BayesRC were more enriched for causal variants than Class I in the Seq_BayesRC (Table 1). The latter included all known NSCoding variants, many of which did not affect milk traits. In Lact_BayesRC, because Class I are more enriched for SNP affecting milk production, a higher proportion of SNP in this class are estimated to belong in the non-zero distributions (Table 1).

For Milk yield we found that the proportion of SNP in the non-zero distribution was enriched for Lact_BayesRC and relative to their small number, they explained a greater proportion of the genetic variance of the trait (Table 1). However, because there are so many SNP in Class III, these SNP still explained most of the total genetic variance of the trait. In Seq_BayesRC there was less enrichment for causal variants than Lact_BayesRC in Class I which comprised only NSCoding, because this allocation

ignored any trait specific biology (although there was still some increase in per SNP variance).

Table 1. Statistics for variants in each class for the Lact_BayesRC and Seq_BayesRC analyses.

Analysis	Statistic	Class I	Class II	Class III
Lact	No. of SNP	4650	64518	924851
Lact	% SNP in non-zero dist	4	1.3	0.8
Lact	% genetic var explained	4	13	85
Seq	No. of SNP	45026	578734	370259
Seq	% SNP in non-zero dist.	0.8	0.7	1.1
Seq	% genetic var explained	6	40	55

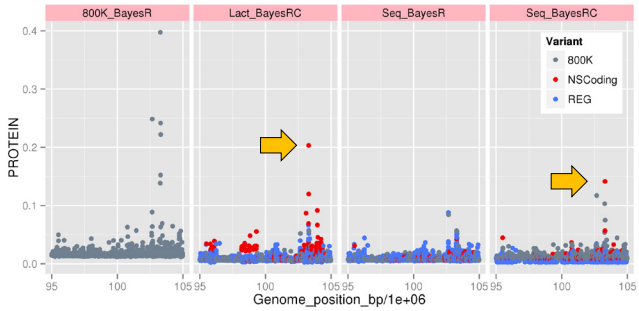

Figure 3. Posterior probabilities for BayesR and BayesRC analyses of Protein on a region of chromosome 11, indicating a NSCoding mutation in the PAEP gene previously associated with protein traits.

Fig. 3 shows the region on Chromosome 11 around the PAEP gene (alias β lactoglobulin). This gene codes for one of the main whey proteins in cows' milk. The highlighted NSCoding SNP in Fig. 3 was previously identified as associated with protein traits (Braunschweig & Leeb 2006), although this SNP may not be the causal mutation. Like DGAT1, this gene was included in the Lact gene set and again it was the Lact_BayesRC analysis which most clearly identified this mutation. The Seq_BayesRC analysis also identifies this mutation but with a lower posterior probability than the Lact_BayesRC. It is of interest to note that while Seq_BayesR struggled to distinguish any one SNP over others in this PAEP gene region, the 800K_BayesR shows a much higher posterior probability for a SNP in very close proximity to PAEP (1636 bp). The reason for this is that in the 800K data there were very few SNP in moderate to high LD with the NSCoding PAEP mutation while in the SEQ data there were over 100 SNP in high LD (r^2>0.75) with the mutation. Therefore the Seq_BayesR analysis did not have any means of distinguishing well among these SNP in high LD. This highlights a dilemma of using sequence data because high LD among very dense variants will sometimes "water-down" the effect of a single causal mutation even when it is present in the data if LD is very high between the QTL and many other SNP in the same region. However the results from the BayesRC analysis demonstrate that good *a priori* biological information can help to overcome this problem when the some classes of SNP are enriched for causal mutations.

Our final example in Figure 4 is a NSCoding variant in the SMEK1 gene, which has not been directly linked to milk traits before but is implicated in energy metabolism and the Insulin/IGF pathway (Yoon et al 2010). SMEK1 was not included in the Lact set, so the NSCoding variant was assigned to Class III in the Lact_BayesRC analysis. None the less this variant shows a high posterior probability as well as a REG SNP (also class III) close by. As would be expected, the highest probability for this SNP is in the Seq_BayesRC where it was assigned to Class I. The 800K_BayesR did not detect any signal in or near this gene, probably because this NSCoding SNP is at a MAF of < 0.01 and is only segregating in the Holsteins.

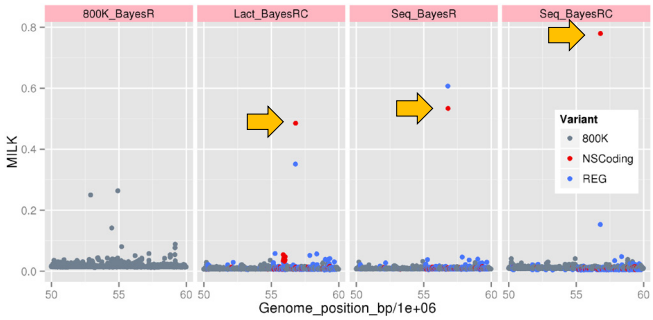

Figure 4. Posterior probabilities for BayesR and BayesRC analyses over a region of chromosome 21, indicating a NSCoding SNP in the SMEK1 gene associated with milk yield.

Conclusion

We expect that the performance of BayesRC would improve if we were able to define a small class or classes that contained a larger proportion of causal variants or SNP in high LD with QTL. Generally however, the BayesRC method shows potential for QTL discovery among sequence variants.

Literature Cited

Braunschweig M.H. & Leeb T. (2006). J Dairy Sci. 89:4414-9.

Erbe M., Hayes B.J., Matukumalli L.K. et al. (2012). J. Dairy Sci. 95:4114-29.

Vander Jagt, C.J. (2012). PhD Thesis. University of Melbourne, Australia.

Wray N.R. (2005). Twin Res. Human Genet. 8:87–94

Yoon Y.S., Lee M.W., Ryu, D. et al. (2010). Proc. Nat. Acad. Sci., 107:17704-17709.