A Research Plan for the Identification of QTL

M.E. Goddard[1,2], I.M. MacLeod[1], K.E Kemper[1] , C.J. Vander Jagt[2], K. Savin[2], C. Schrooten[4] and B.J. Hayes[2,3].
[1]University of Melbourne, [2]Department of Environment and Primary Industries, [3]Latrobe University, Melbourne, Australia,
[4]CRV, 6800 AL, Arnhem, Netherlands.

ABSTRACT: Past attempts to find mutations causing variation in quantitative traits important in livestock production have largely been thwarted by the small effects that individual mutations have. However, we now have more powerful tools, such as genome sequence on individual animals, that increase our power to identify the mutations underlying quantitative trait loci (QTL). Here we review the statistical methods used to analyze association studies and the biological information that might be used to discriminate between alternative sites as the causal mutation. As well as the small effect of most QTL, it is also difficult to identify causal mutations because they are typically in LD with other polymorphic sites and so one cannot decide which is the causal mutation. The problem caused by LD can be reduced by using multiple breeds and by fitting all sequence variants simultaneously with a Bayesian model in which many variants are expected to have no effect. The effect size of a QTL may be increased by using traits, such as gene expression, that are close to the primary effect of the mutations. Biological information, such as that provided for the human genome by the ENCODE project, can be incorporated into the statistical analysis of the association study so that objective estimation of the probability that each sequence variant has an effect on the trait can be made.
Keywords: QTL, ENCODE, gene expression, milk

Introduction

The tools of modern genomics are sufficiently powerful that many mutations causing simply inherited Mendelian traits have been identified. However, the variation in complex or quantitative traits is caused by mutations at many sites in the genome and by non-genetic or environmental variables and we have had much less success in identifying the mutations that cause variation in quantitative traits (the QTL). Ideally we should like to know the gene in which the mutation occurred (or the gene whose regulation is changed by mutations in non-coding DNA) and the specific site that has been mutated causing a change in the trait. This has been achieved in some cases (eg DGAT, Grisart et al 2004) but they tend to be QTL with atypically large effects. In most cases the identified QTL explain only a small proportion of the genetic variance in a quantitative trait. Nevertheless, it can be expected that with better data and more powerful tools we can gradually identify QTL of smaller and smaller effect and which collectively explain more and more of the genetic variance. In this paper we review methods that may help us achieve that aim.

Prior to the availability of "SNP chips" our ability to identify QTL was limited by the low density of markers that could be used. Usually linkage mapping was used to map a QTL but often the confidence interval for its position was as large as 50 cM. Then, progressively finer mapping, using markers newly discovered in the target region, was used to reduce the confidence interval. This fine scale mapping relied on linkage disequilibrium (LD) between the causal mutation and the markers. Often a biologically sensible guess was made as to the relevant gene and sequencing carried out to find possible mutations. The availability of whole genome sequencing at an affordable price should greatly speed up this process because the causal mutations should be included in the genome sequence data. By using a database of animals with genome sequence, it is now possible to impute full genome sequence into animals with SNP genotypes and phenotypes and carry out an association analysis to find the QTL without any preliminary mapping experiments (eg Daetwyler et al. 2014). However, it will still be very challenging to identify mutations if they cause only a small proportion of the variance in a trait, and so it will be necessary to use a number of tools and approaches to achieve the aim. In this review we divide the approaches into statistical analysis of association between polymorphisms and phenotype, and use of biological information that is external to the association study. Before considering these two classes of tool, we discuss the advantages of identifying the causal mutations or QTL.

Advantages of Identifying QTL

Identification of QTL would have benefits in basic science and in genomic selection. Our understanding of quantitative traits and their evolution would be greatly improved if we could study them at the level of causal mutations. If we knew the gene underlying a QTL, even if not the site of mutation, we would better understand the physiology involved and might be able to intervene, for instance, using a pharmaceutical. For selection we need to know the actual polymorphic site in the genome or at least a site in high linkage disequilibrium (LD) with it.

How much would identification of QTL increase the accuracy of genomic selection? Meuwissen et al (2001) introduced 3 methods to predict the breeding value (BV) of animals from genotypes at a panel of dense markers covering the genome. Within a breed, such as Holstein cattle, the method they called "BLUP" can predict BV with accuracy > 0.8 (Wiggans et al 2011). This appears to leave little room for increased accuracy from identifying causal mutations. However, no other breed can match the quantity and quality of data available to train the prediction equation that is available for milk yield in Holstein, so it will not be easy for other breeds and species to match this accuracy. Even in Holsteins the SNPs explain only 32-80% of the genetic variance in a trait (Haile-Mariam et al 2013)

and this limits the accuracy of genomic EBVs . In addition the prediction equations are not robust. For instance, a prediction derived from Australian Holsteins has a low accuracy when tested in NZ Holsteins (Pryce et al 2012). Some traits, such as feed conversion efficiency, are so expensive to measure that we need to combine data across breeds, so methods that do this would be advantageous.

An alternative to a within breed prediction equation based on medium density markers is a prediction combining data from multiple breeds based on dense markers and using a Bayesian method such as Bayes R (Erbe et al 2012, Kemper et al these proceedings). In simulation studies Bayesian methods give higher accuracy and greater robustness than BLUP in populations with high effective population size (Ne) and use of causal mutations further increases accuracy compared to dense markers (Meuwissen and Goddard 2010). We tested this in real dairy cattle data by comparing BayesR prediction accuracy using either HD 800K SNP or a larger set of imputed sequence SNP (for a more detailed description see Hayes et al, this proceedings). The latter included all predicted non-synonymous coding SNP because these are more likely to be causal mutations (Table 1). We found a consistent advantage for fat, protein and milk yield using the imputed sequence data. More evidence from real data is needed but it seems likely that identification of causal mutations would increase accuracy of predicting BV in many situations.

Table 1. Correlations between predicted GEVB and de-regressed proofs (derived from MACE EBV) using either HD 800K SNP or a subset of imputed sequence variants (Seq subset).

SNP set	Fat Yield[1]	Prot Yield[1]	Milk Yield[1]
800K	52.6%	58.6%	53.3%
Seq subset	54.2%	60.5%	54.8%

[1]Predictions equations were calculated from a combined breed reference (Australian Holstein and Jersey) with DTD and were validated in Red Holsteins.

Analysis of Association Studies

Statistical models. Genome wide association studies (GWAS) are commonly used for two purposes – to map causal mutations and to derive a prediction of BV for use in genomic selection. For the former purpose, the data are usually analyzed one SNP at a time and the SNPs with the strongest association with the trait reported. For the latter purpose, all SNPs are fitted in a single prediction equation. It would be better to fit all SNPs simultaneously regardless of which was the purpose of the experiment. This decreases false positives due to population structure and increases power. Population structure, such as admixture of two different populations, can cause false positives because it can cause an association between a marker and a QTL even if they are not linked. However, more subtle forms of structure, such as half-sib families, also generate more false positives than expected (MacLeod et al 2010). To avoid this problem one can fit the effect of the structure (eg by fitting a polygenic term described by the pedigree) or one can fit markers that adequately cover the whole genome simultaneously. This increases power because some of the error variance is explained by the other markers so the effect of any given marker is estimated with less error (Yang et al 2014).

To fit all SNPs simultaneously they should be treated as random effects as they are in the prediction equations of genomic selection. The various methods suggested for genomic selection vary in the distribution they assume for the effects of markers. When analyzing sequence data that may contain 30 million polymorphic sites, it would be logical to assume a prior distribution in which many variants have no effect. Methods called Bayes B (Meuwissen et al 2001), Bayes Cpi (Habier et al 2011), Bayes R (Erbe et al 2012) do this but BSLMM fits a small effect of all markers as well as larger effects of a few markers (Zhou et al 2013). These methods report exactly the statistic required for identification of QTL, that is, they report the posterior probability conditional on the data that a polymorphic site affects the trait.

A commonly used analysis in human genetics (the so-called mixed linear model) fits one SNP as a fixed effect and all SNPs as random effects under the BLUP model where all their effects are drawn from a normal distribution. This model sacrifices some power because the effect of the target SNP is fitted twice, once as a fixed effect and once as a random effect. (Yu et al 2006 and reviewed by Yang et al 2014).

Power to find causal mutations. To detect a QTL its effect must be larger than the standard error. This demands large sample sizes or QTL of atypically large effect. For instance, a typical QTL, that explains 0.001 of the phenotypic variance, has a probability of 0.5 to reach a p-value of $5*10^{-8}$ if the sample size is 30,000. Even with such large sample sizes, it is unlikely that the causal mutation will be the most significant SNP in the vicinity or that it will have the highest posterior probability of affecting the trait. Each causal mutation will have many SNPs in LD with it. For instance, in Holstein a typical non-synonymous coding SNP has LD r^2 >0.75 with an average of ~15 other SNPs spanning ~150Kb. Sampling error in the estimated effects of these SNPs will probably cause one of them to have a stronger association with the phenotype that the causal mutation. Therefore, it is necessary to use other sources of information to attempt to isolate the correct causal mutation or at least a marker close to it and in near complete LD with it. The remainder of the review considers such additional sources of information. Even if we do not pin point the correct mutation, it would be very useful to know the size of the confidence interval in which the QTL is located. Fernando (these proceedings) considers this.

Multiple breed data. The LD between SNPs and QTL is decreased when two or more breeds are combined. This is a disadvantage if only 50k markers are used (de Roos et al 2008) but it is an advantage when using dense markers or sequence. Even if the QTL only segregates in one breed, the multi-breed data is an advantage because markers that segregate in more than one breed can be ruled out. Our experience is that some QTL segregate in more than one breed and some do not but few segregate in both Bos indicus and B. taurus (Bolormaa et al 2013). Raven et al (2014) found evidence that only ~15% of milk yield QTL

were shared between Holstein and Jersey although this was also affected by the power of the analysis.

Multi-trait analysis. QTL often affect more than one trait. When separate analyses are carried out for each trait the most significant SNP may vary from trait to trait by chance. By combining information across traits a multi-trait analysis has increased power to identify the QTL (Bolormaa et al 2014).

The ability to detect, map and identify a QTL is increased if the QTL explains a large amount of the variance of a trait. This suggest that power would be increased by including traits that are closer to the direct effect of the QTL. However, attempts to exploit this idea, for instance by using physiological traits such as hormone concentrations, have not often been successful. One particularly useful trait is gene expression, that is the concentration of mRNA for a particular gene in the cells of a given tissue. QTL for gene expression (eQTL) fall into two groups – cis eQTL are mutations that affect expression of the gene on the same DNA molecule or chromosome and trans eQTL affect expression of both copies of the gene. In practice, any eQTL that maps near to the gene is called a cis eQTL. Cis eQTL often explain a large part of the variance in gene expression (Visscher et al 2012) and so we have higher power to map and identify them than is the case with most QTL. Cis eQTL have two other advantages – they immediately identify the gene involved and they show allele specific expression. Allele specific expression means that one of the two alleles of a gene is expressed more highly than the other. This can be detected by RNA sequencing. This is a very sensitive measure of an eQTL because it is a within-animal comparison in which most variables that affect expression are held constant. If an eQTL is found it does not prove that the same mutation also causes variation in a conventional phenotype. Evidence for this is that the two traits (gene expression and conventional phenotype) segregate together.

Use of Biological information

There is usually some information outside the association study that would help in deciding which polymorphic sites affect the trait as opposed to being associated with the trait due to LD. This outside information might be about the probability that a particular gene affects the trait or the probability that a particular site in the genome affects the trait. If the information about eQTL discussed above was external to the association study it would fall in this category and could contain information about both the gene and, possibly, the specific site.

Physiological knowledge of a trait may indicate that some genes are more likely than others to affect the phenotype. This information might be used before the experiment is carried out in a candidate gene approach. Alternatively, the genes in the confidence interval of a QTL may be considered and prioritized based on their known involvement with the trait. Unfortunately, a case can be made for almost any gene influencing any trait so such post hoc arguments are not always convincing. Another approach is to list all the genes within the confidence interval of any QTL throughout the genome and ask which pathways are over represented in the list (Weng et al 2011).

Instead of focusing on known candidate genes, one could focus on 'candidate sites'. For instance, one could examine the effect on phenotype of all mutations that generate a premature stop codon. This is an example of "reverse genetics" in which one progresses from the mutation to the phenotype instead of the conventional path of starting with an inherited phenotype and looking for the mutations that causes it.

Some sites in some genes are more likely to affect phenotype if mutated than others. Typically sites that change an amino acid in a protein (non-synonymous coding sites) are thought to have more affect than synonymous sites or non-coding sites. However, although many Mendelian mutations with a large effect are coding mutations, many QTL appear to be non-coding mutations. In human genetics the ENCODE project identified many sites in the genome outside gene coding regions that may be functional and hence display a phenotype if mutated (Djebali et al 2012). For instance, these sites include transcription factor binding sites and sites with histone marks indicating "open" chromatin. It would be very valuable to have a catalogue of eQTL and other sites in livestock genomes that appear to be functional.

Another indicator of a site where mutation will generate a phenotypic effect is the conservation of the site across species. About 5% of the genome is conserved across mammals suggesting that mutations in these regions would be detrimental. However, Kindt et al (2013) and Koufariotis et al (2014) did not find conserved sites enriched for QTL for human and bovine traits respectively.

It is desirable to integrate outside biological knowledge into the framework of the association study so that more objective decisions can be made about the probability that a particular mutation affects the trait. One way to do this is to define categories of sites and to estimate the proportion of sites in each category that affect the trait. We use a variant of Bayes R called Bayes RC to do this (MacLeod et al, these proceedings). For instance, sites might be classified into 2 categories (non-synonymous and others) and the proportion of sites that affect a trait estimated in each category. This uses the known categorization of sites without any prejudice about which categories are important to a particular trait.

Evidence of selection

Long term selection changes allele frequency at a selected site and at neighboring sites. Thus comparing populations subject to different past selection pressures might help identify genome sites that affect the selected traits. The power of this approach is reduced by linkage and by drift in gene frequencies due to finite population size (Kemper et al 2014).

One generation of selection can affect genotype frequencies if it is very strong. For instance, a mutation that is a recessive in utero lethal will lead to offspring whose genotype frequencies depart from Hardy-Weinberg equilibrium due to the missing homozygote class. This has been used to identify mutations causing in utero mortality (Van Raden et al 2011).

Some examples

Using 16,000 Jersey and Holstein bulls and cows (dataset 1) with phenotypic or progeny test records for milk production traits (DTD), we carried out a GWAS using Illumina high density or 800k SNPs which were either genotyped or imputed from lower density SNP genotypes. The single SNP regression results for milk on chromosome 6 are shown in Figure 1 as well as a BayesR analysis of the same data. There is a small peak around 87 Mb where the casein cluster is located but this extends across to a larger peak at ~89 Mb. The Bayes R analysis shows these as 2 clearly separate peaks: one above the casein gene cluster (red bar) and one close to the GC gene (purple bar). Using the '1000 bull genomes' database we imputed all SNPs within 2 kb of genes and carried out a new Bayes R analysis (Figure 2). This analyses was carried out in a set of 8000 Holstein and Jersey bulls (dataset 2). The Bayes R analysis with sequence gives a high and almost equal posterior probability for 2 SNP at ~ 89 Mb while the casein cluster shows a smaller peak with several SNP all showing similar posterior probabilities.

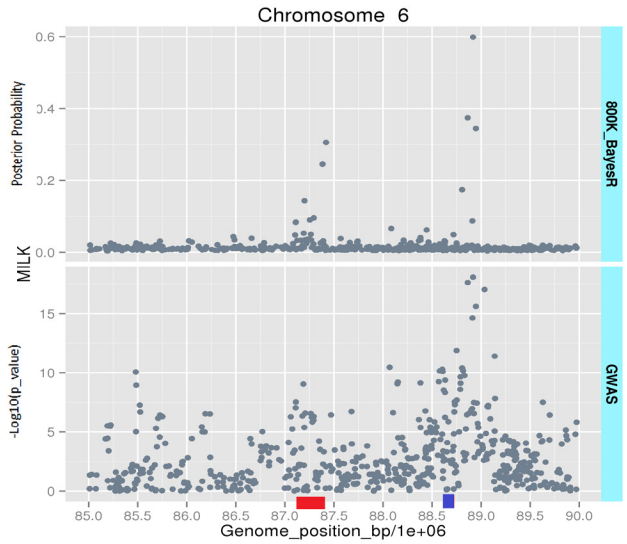

Figure 1. Comparison of results from a GWAS 800K analyses for Milk Yield and a BayesR analyses using the same data. The red bar shows the casein gene cluster and the purple bar the GC gene.

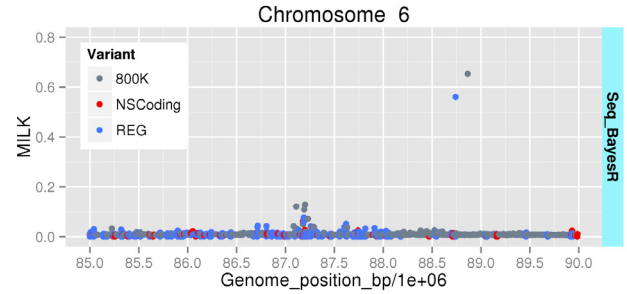

Figure 2. Posterior probabilities for BayesR analysis of Milk Yield using imputed sequence in the same region of BTA6 as shown in Figure 1.

Vander Jagt (2012) analyzed a series of microarray experiments on gene expression in the mammary gland of lactating cows. She listed all genes that were differentially expressed in two or more experiments. The SNPs in or within 50 kb of these genes were identified in the genome sequence data. Using this independent biological information we carried out a further analysis of dataset 2 using the Bayes RC method, described briefly above. The SNPs were classified into three groups; (1) non-synonymous SNPs in genes listed by Vander Jagt (2012), (2) other SNPs in or near these genes, (3) all other SNPs. Bayes RC estimates the proportion of SNPs with non-zero effects on milk yield separately for each of the 3 groups. We found A higher proportion of non-zero effects in group 1. The analysis also reports the posterior probability that each SNP has an effect on milk yield (Figure 3, Lact_BayesRC). This shows that there is a non-synonymous SNP in beta casein (CSN2) with a posterior probability of 0.35 of affecting milk yield (higher than for the Bayes R analysis, Figure 2) and a group 2 SNP (non-coding near a differentially expressed gene) with the highest posterior probability in this region of 0.8. This second SNP is at 88.741762 Mb, 2564 bp upstream from the gene GC located at 88.69 – 88.74 Mb. Bayes R, without the biological grouping of SNPs, does detect this SNP but with a low posterior probability of affecting milk yield due to very high LD between many SNP in this region. The SNP segregates in Holsteins (MAF = 0.42) but is almost fixed in jerseys (MAF=0.004) which is in accord with the detection of the QTL only in Holsteins.

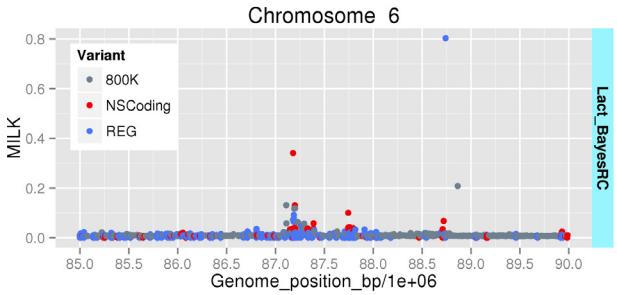

Figure 3. Posterior probabilities for BayesRC analysis of Milk Yield (Lact_BayesRC) with imputed sequence in the same region of BTA6 as shown in Figure 1 and 2. Prior to analysis the SNP were assigned to 3 different classes based on lactation biology and genome information.

We also carried out a GWAS using data on average milk yield and average allele frequencies from 12 different cattle breeds (dairy and beef). Twenty two SNPs near GC (from 88.728 – 88.751 Mb) were significantly associated with milk yield ($p<0.001$). The most significant SNP ($p< 0.00001$) is at 88.743767 Mb, less than 2 kb from the high posterior probability SNP of Bayes RC.

The GC gene (Group-specific Component, Chr6:88695939..88739180) encodes the vitamin D binding protein (VDB), the main transporter of vitamin D in plasma. It is expressed mainly in liver (Haddad, 1979) but

also in other tissues. Vander Jagt (2012) found that in mammary gland, GC mRNA levels are up-regulated 1.5 fold in cows treated with growth hormone and down regulated ~1.6 fold in an in vitro experiment when cells were grown on ECM compared to plastic with no prolactin in the media in both cases.

Further evidence for a possible role for vitamin D and its carrier protein encoded by GC in regulating milk yield come from our understanding of vitamin D metabolism and its role in the mammary gland. Via the plasma, VDB carries the sterol vitamin D3 from its site of synthesis in the skin to the liver, its 25(OH)D3 derivative from the liver to the kidney and finally the active form, $1,25(OH)_2D3$, from the kidney to various other tissues (reviewed by Omdahl et al, 2002). One of the target tissues is the mammary gland. In tissue culture $1,25(OH)_2D3$, via the vitamin D receptor, restricts growth and differentiation of mammary epithelial cells (Zinser et al 2002; Zinser & Walsh, 2004; Kemmis et al, 2006). Milk production depends on the number and activity of these cells. Raven et al (2014) also found that a SNP in the gene for vitamin D receptor (Chr5:32550521..32606144) was associated with milk yield.

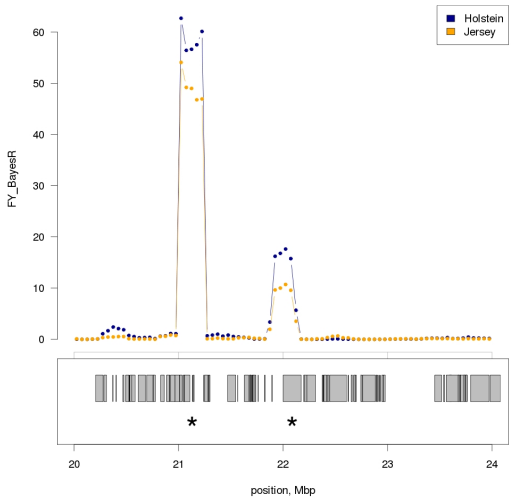

Figure 4. Variance explained by 250 kb windows on chromosome 26 for fat yield. * mark the position of genes SCD and BTRC. The vertical scale is the variance explained by each window relative to the mean of all windows.

Kemper et al (these proceedings) used Bayes R analysis of milk production traits to distinguish 2 QTL on chromosome 26 between 21 and 22.5 Mb. They calculated the variance in GEBV for fat yield for 250 kb windows based on BayesR estimates of SNP effects (dataset 1). One QTL is near the gene stearoyl co-enzyme A desaturase (SCD) and one near FGF8 and BTRC (Figure 4). Figure 5 shows the posterior probabilities for individual SNPs in imputed sequence data (dataset 1). No single SNP near BTRC has a high posterior probability of affecting fat yield because many SNPs all have a low probability indicating that the analysis could not distinguish which of them was causal. Near SCD in the BayesR analysis of the 800k SNP data there is a single SNP with posterior probability of 0.49. However, in the sequence data no single SNP reaches this probability because the analysis cannot distinguish among the many SNPs in high LD (Seq_BayesR). However, in the Bayes RC analysis (Seq_BayesRC, details in MacLeod et al. these proceedings) a non-synonymous coding SNP in SCD has a posterior probability of 0.6. The SCD mutation is segregating in both Holstein (MAF 0.3) and in Jersey (MAF 0.2). SCD plays an important role in fatty acid desaturation and this same SCD mutation was shown to be associated with fatty acid composition in Holstein milk (Mele et al 2007). Also Rincon et al (2012) identified another synonymous SNP associated with fatty acid composition in complete LD with this one.

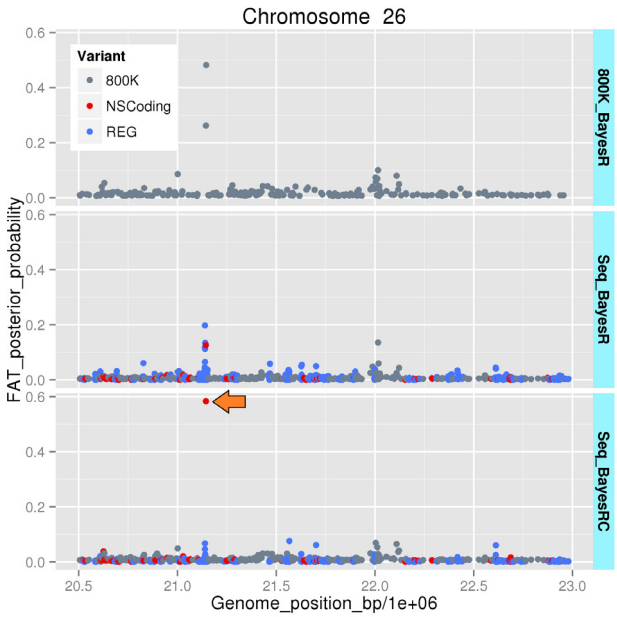

Figure 5. Posterior probabilities for 800K BayesR analysis, contrasted with imputed sequence BayesR and BayesRC analyses for fat yield. In BayesRC analysis prior knowledge of the variant type was used to assign SNP to 3 classes. The arrow highlights a SNP in the SCD gene.

These examples illustrate two statistics derived from BayesR analyses – the variance explained by a window and the posterior probability that individual SNPs affect the trait. Both are useful. The window variance may detect a QTL more easily than posterior probabilities because the no single SNP may have a high probability. However, a high posterior probability when all sequence variants have been included in the model, is evidence that the causal mutation or one in very high LD with the causal mutation has been found.

Conclusion

There would be advantages to basic science and to selection of livestock if mutations causing variation in quantitative traits could be identified. However, most of the variance in quantitative traits is caused by numerous mutations each of small effect. To find at least the more important of these, we advocate analysis of multiple traits and genome sequence data on large samples from multiple

breeds using Bayesian models that fit all sequence variants simultaneously. Despite the power of such a design, due to LD there will still be many polymorphisms that could be the cause of each QTL. To discriminate between these, biological information about the possible genes and genomic sites can be used but we advocate an objective method of incorporating this information. We illustrated these recommendations by identifying SNPs near the gene for vitamin D binding protein as affecting milk yield possibly through regulation of the expression of this gene (GC).

Literature Cited

Bolormaa S, Pryce JE, Kemper KE et al. (2013). Genet Sel Evol. 45:43.

Bolormaa S, Pryce JE, Reverter A et al. (2014). PLoS Genet. (Accepted)

de Roos APW, Hayes BJ, Spelman RJ, Goddard ME. (2008). Genetics. 179:1503-1512.

Daetwyler et al. (2014). Nature Genetics. Accepted.

Djebali S, Davis CA, Merkel A, et al. (2012). Nature. 489:101-108.

Grisart B, Farnir F, Karim L. et al. (2004). Proc. Nat. Acad. Sci. 101:2398-2403.

Haddad JG Jr. (1979). Clin Orthop Relat Res. 142:249-61.

Haile-Mariam M, Nieuwhof GJ, Beard KT et al. (2013). J. Anim Breed Gen. 130:20-31.

Kemmis CM, Salvador SM, Smith KM, Welsh J (2006). J Nutr. 136(4):887-92.

Kemper KE, Saxton SJ, Bolormaa S et al (2014). BMC Genomics. (submitted).

Kindt AS, Navarro P, Semple CA and Haley CS (2013) BMC Genomics. 14:10.

Koufariotis et al. (2014) BMC Genomics. Submitted

MacLeod IM, Hayes BJ, Savin KW et al. (2010). J Anim Breed Genet. 127:133-142.

Maurano MT, Humbert R, Rynes E, et al. (2012). Science. 337:1190-1195.

Mele M, et al. (2009). J Dairy Sci. 92(1): 392-400.

Meuwissen, THE, Hayes BJ and Goddard ME (2001). Genetics. 157:1819-1829.

Omdahl JL, Morris HA, May BK (2002). Annu Rev Nutr. 22:139-66.

Pryce JE, Arias J, Bowman PJ et al. (2012). J Dairy Sci. 95:2108-2119.

Raven L-A, Cocks B and Hayes BJ (2014). BMC Genomics. 15:62.

Rincon, Gonzalo, et al. (2012). J Dairy Research 79:66-75.

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

VanRaden, P.M. Olson, K.M., Null, D.J. et al. (2011). J Dairy Sci. 94:6153–6161.

Weng L, Macciardi F, Subramanian A, et al. (2011). BMC Bioinformatics. 12:99.

Kindt AS, Navarro P, Semple CA, and Haley CS (2013). BMC Genomics. 14:10.

Wiggans GR, VanRaden PM, and Cooper TA (2011). J Dairy Sci. 94:3202-3211.

Yang J et al (2014). Nat Genet. 46:100-106.

Yu J et al. (2006). Nat. Genet. 38, 203–208 (2006).

Zhou X, Carbonetto P, Stephens M (2013). PLoS Genet. 9:e1003264.

Zinser G, Packman K, Welsh J (2002). Development. 129(13):3067-76.

Zinser GM, Welsh J (2004). Mol Endocrinol. 18(9):2208-23.