Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture

M. E. Goddard1,2, K. E. Kemper3, I. M. MacLeod1,2,3, A. J. Chamberlain2 and B. J. Hayes2,4

1Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Melbourne, Victoria 3010, Australia
2Department of Economic Development, Jobs, Transport and Resources, and 3Dairy Futures Cooperative Research Centre, Agribio, La Trobe University, Bundoora, Victoria 3083, Australia
4School of Applied System Biology, La Trobe University, Agribio Building, Bundoora, Australia

Complex or quantitative traits are important in medicine, agriculture and evolution, yet, until recently, few of the polymorphisms that cause variation in these traits were known. Genome-wide association studies (GWAS), based on the ability to assay thousands of single nucleotide polymorphisms (SNPs), have revolutionized our understanding of the genetics of complex traits. We advocate the analysis of GWAS data by a statistical method that fits all SNP effects simultaneously, assuming that these effects are drawn from a prior distribution. We illustrate how this method can be used to predict future phenotypes, to map and identify the causal mutations, and to study the genetic architecture of complex traits. The genetic architecture of complex traits is even more complex than previously thought: in almost every trait studied there are thousands of polymorphisms that explain genetic variation. Methods of predicting future phenotypes, collectively known as genomic selection or genomic prediction, have been widely adopted in livestock and crop breeding, leading to increased rates of genetic improvement.

1. Introduction

Complex or quantitative traits are important in medicine (e.g. diabetes), agriculture (e.g. yield of rice) and evolution (e.g. body size). These traits are called complex because they are controlled by many genes and by environmental factors. Although the genetics of quantitative traits has been studied for over 100 years, very few of the polymorphisms that cause variation in these traits were known until recently. The development of assays that could determine the genotype of an individual at thousands of single nucleotide polymorphisms (SNPs) has revolutionized the study of complex traits. The SNPs assayed might be neutral polymorphisms with no effect on the traits studied, but linkage disequilibrium (LD) between the SNPs and the causal polymorphisms generates an association between the traits and some of the SNPs. Genome-wide association studies (GWAS), which assay a genome-wide panel of SNPs, have discovered thousands of associations between SNPs and complex traits [1]. These GWAS are intended to map the causal polymorphisms to a region of the genome but do not identify them.

The data from GWAS can be used for three purposes. First, they can be used to predict future phenotypes. In human medicine, this might be the probability that a person will develop type 2 diabetes in the future. In agriculture, it is usually the phenotype of the offspring of animals or plants that we wish to predict so that those with the best breeding value can be selected as parents of the next generation. Although the SNPs used may have no causal relationship with the trait, they may still be useful for prediction due to their LD with causal variants. Second, GWAS data are used to map the causal variants to a region of the genome and hopefully to identify them. This increases our understanding of the biology of complex traits and may suggest methods of controlling them such as new drug targets. Third, GWAS data provide an overview of...
the genetic architecture of complex traits that is useful in medicine, agriculture and evolution. We would like to know how many polymorphisms control a trait, what are their effects and allele frequencies, the LD between them, and how they evolve.

Usually different methods of analysis of GWAS data have been used for each of these three purposes. For instance, mapping causal polymorphisms is usually done by fitting one SNP at a time in a regression model [2]. Conversely, predicting genetic value is most often done by assuming all SNPs have an effect drawn from the same normal distribution [3]. In this paper, we will argue that a new model, where SNPs are assumed to have an effect drawn from a mixture of normal distributions with increasing variances, can be used for genomic prediction, mapping of causal variants and inference on the genetic architecture of complex traits. The rest of the paper begins by introducing the statistical model and its use for prediction of phenotype. This is followed by describing the use of the same model for mapping of causal variants and understanding the genetic architecture of complex traits. We then discuss the limitations of the method, and the implications for our understanding of complex traits and for future research on prediction of phenotype and mapping of causal variants.

2. Prediction of genetic value

In many datasets, the number of SNPs (p) is greater than the number of individuals with records (n). Consequently, if the effects of the SNPs on the trait are treated as fixed effects in a multiple regression analysis, there is no unique solution. However, the total variance explained by all SNPs (a result of their effect sizes and allele frequencies) must be less than the total genetic variance, and this places a restriction on the effect sizes. More accurate predictions can be obtained by treating the effect sizes as random variables drawn from a distribution which is consistent with the total genetic variance [4]. In general, the best prediction of a random variable (g) from a set of predictors (x) is $E(g|x)$, that is, the expected or average value of g conditional on the values observed for x [S]. (Here best means minimum mean-squared errors.) We will restrict discussion to a linear predictor of the form $b'y$, where b is a vector of regression coefficients. Then the best prediction rule implies that we estimate b by $E(b|y,x)$, where data might be the genotypes (x) and phenotypes (y) of individuals from a GWAS. In this formulation of the problem, the elements of b (h) are the apparent effect of the SNP on the trait and are treated as random effects drawn from a distribution (p(h)). As there are typically thousands of SNPs, it is possible to imagine the distribution of SNP effects on a trait as simply the distribution of these thousands of effects: many may have no effect at all and some may have a large effect.

Therefore, $E(b|y,x) = \int p(b|y,x)\,db$, which can be re-expressed using Bayes theorem as $\int \int p(b|y,x)\,db\,dy = \int \int p(y|b)\,p(b|y)\,db$ for all y. Here, $p(y|b,x)$ is the likelihood of the phenotypes (y) given the genotypes of the individual (x) and the effects of the SNPs (b). This is the method invented by Meeuwissen et al. [4] and called genomic selection or genomic prediction.

The statistical analysis resulting from applying this best prediction rule depends on the prior distribution chosen for b. Meeuwissen and colleagues considered three possible prior distributions [4]. In one, b is assumed to be normally and independently distributed with a mean of 0 and a variance ($\sigma^2$) that is same for all SNPs; this method is an example of best linear unbiased prediction (BLUP). In the other two, Meeuwissen et al. used a t distribution, and a mixture of zero and a t distribution. The method called ‘Bayes R’ by Erbe et al. [6] is a further development, which uses a prior distribution for b which is a mixture of four normal distributions each with zero mean but with variances of 0, 0.0001 $\sigma^2_1$, 0.001 $\sigma^2_2$ and 0.01 $\sigma^2_3$. The mixing proportions are estimated from the data, so this is a flexible prior that can approximate many possible distributions for b (the SNP effects). In our applications of this model, we have assumed that the mixing proportions are drawn from a Dirichlet distribution with parameters (1, 1, 1, 1). This is a deliberately vague prior so that it has little impact on the final estimates of the mixing proportions which are driven mainly by the data.

While BLUP is a linear method in that b is estimated by a linear combination of the phenotypic data y, the other methods are nonlinear in y. In this paper, we advocate the use of these nonlinear models with particular reference to Bayes R.

The BLUP prior corresponds to a ‘pseudo-infinitiesimal’ model in which all polymorphic sites in the genome have an effect on every trait, and all effects are of similar magnitude and very small. For instance, if there are 1 million SNPs, each one is assumed to explain approximately $10^{-5}$ of the genetic variance ($\sigma^2_g$). As a consequence of this assumption, all estimated SNP effects are shrunk severely towards 0 when BLUP is used. The other models allow the distribution of SNP effects to depart from this pseudo-infinitiesimal distribution, with some SNPs having zero effect and some SNPs having a large effect on the trait.

Genomic prediction requires a reference or training population in which the individuals have both phenotypes and genotypes. Analysis of these data generates a prediction equation which can then be used to predict genetic value in individuals with genotypes but without phenotypes. In accordance with convention in genetic evaluation, if not in statistics, we will define the ‘accuracy’ of the prediction as the correlation between the predicted genetic value and the true genetic value among these individuals. The factors determining the accuracy when BLUP is used have been considered in theory by Daetwyler et al. [7] and Goddard [8]. The accuracy of predicting genetic values depends on the proportion of the genetic variance explained by the markers (referred to below as SNPs) and the accuracy with which the effect of those SNPs is estimated. Both components of accuracy depend on the LD within the genome. Low LD increases the number of individuals with records and the number of SNPs needed to achieve a given accuracy. Consequently, accuracy is typically lower in humans than within a breed of cattle, where long-distance LD exists due to small recent effective population size. In cattle, the proportion of genetic variance explained by SNPs is in the range 0.5–0.9, and in humans it is 0.3–0.5 for many traits [9–12].

In practice, the accuracy of genomic prediction using the nonlinear methods is equal to or higher than the accuracy of BLUP [12–14]. For example Kemper et al. [14] found a 5% increase in accuracy of genomic prediction for milk yield traits in dairy cattle using Bayes R compared with BLUP, and Moser et al. [12] found an increase in accuracy of genomic predictions of Bayes R over BLUP for Crohn’s disease, rheumatoid arthritis, and type 1 diabetes, but not for bipolar disorder, coronary artery disease, hypertension or type 2 diabetes.
However, the small advantage of nonlinear models over BLUP points to the high number of causal variants affecting most traits.

Genomic selection is now widely used in livestock (especially dairy cattle) and crops. It should double the rate of genetic improvement in dairy cattle [15].

3. Mapping and identification of causal polymorphisms

To map genes for a quantitative trait (QTL, or quantitative trait loci) to a position on the genome using GWAS data, the most common analysis is to fit one SNP at a time in a regression model [1]. While this is straightforward and computationally undemanding, there are several disadvantages to this approach. Many SNPs are likely to be in LD with a single QTL generating many SNPs associated with the trait. Common practice is to focus on the most highly associated SNP within a genomic region (e.g. 2 Mb). However, in livestock and many crops LD extends for a long distance, and so SNPs more than 2 Mb from the QTL may still show a significant association with the trait. A further complication is that there are typically so many QTL for each complex trait that an SNP may be in LD with more than one QTL. Therefore, it is difficult to tell how many QTL are indicated by the GWAS results.

A solution to this problem would be to fit all SNPs simultaneously. In this way, only the SNPs that are necessary to track each QTL should be included in the final model. Usually, there are more SNPs than subjects so there is no unique solution if the SNP effects are treated as fixed effects. A widely used alternative is to fit all SNP effects as random effects using the BLUP model discussed above and then to fit one SNP as a fixed effect [16–19]. However, this does not eliminate the problem and seems illogical—why fit one as fixed and all the rest as random effects? A better solution is to fit all SNP simultaneously as random effects, which is the same model as used for genomic prediction. However, the BLUP model estimates small effects for all SNPs and so does a poor job of mapping QTL. By contrast, the Bayes R model, in which many SNPs have no effect, gives a large effect to SNPs that best track the QTL.

Figure 1 compares partial results of a genome-wide analysis of fat percentage in bovine milk in a region around the FASN gene (data described in [20]), using either Bayes R or the common GWAS method (SNP fitted singly as a fixed effect using the linear mixed model). In the Bayes R model (figure 1a), a single SNP just upstream of the FASN gene has the highest posterior probability, while in the GWAS model (figure 1b) there are several SNP extending across to the CCDC57 (coiled-coil domain containing 57) gene region with almost equally significant effects. FASN has previously been suggested as a candidate gene affecting fat percentage of milk because FASN is a key enzyme in de novo fatty acid biosynthesis (e.g. [21,22]). In an RNAseq study, FASN was also found to be more highly expressed in lactating bovine mammary tissue than in 17 other bovine tissues [23]. Although there may be a second QTL effect
due to a variant of the CCDC57 gene, there is less evidence of this from the Bayes R analysis which fits all SNPs simultaneously, while the evidence from the GWAS single SNP regression predicts almost equal significance with the variants close to FASN. If there are multiple SNPs in high LD, the analysis may be unable to say which is the best one to include in the model. Then all these SNPs may receive a low posterior probability of a non-zero effect. However, when the results are combined across the SNPs, it is clear that a QTL resides in this region.

One way to do this is to predict the genetic value of each individual based only on the SNPs in the region. If this local estimated genetic value has a high variance, it indicates a QTL in this region. An example is given in figure 2. It can be seen that Bayes R gives a sharper position for the QTL than BLUP.

By applying this Bayes R analysis to genome sequence data, rather than data from a panel of SNPs, we would hope to identify the causal polymorphisms directly. Unfortunately, if there are several variants in high LD, the analysis
cannot tell which of them is causal. Also there is sampling error, such that the most significant variant or the one with the highest posterior probability may not be the causative mutation, particularly if the reference population is small. It is helpful in this situation to have independent information on the likelihood that a mutation in each site would affect the phenotype. Typically, in the analysis of GWAS, this information is used after the statistical analysis and in a subjective manner. We have modified Bayes R to include a priori information about the polymorphic sites (Bayes RC [20]). The sites are placed in categories and the mixing proportions for each category are estimated in the analysis. For instance, non-synonymous coding sites may be placed in one category and all other sites in another category. This provides objective evidence for the probability that coding sites are more likely to alter phenotype than non-coding sites. The Bayes RC approach can improve both the precision of QTL mapping and accuracy of genomic prediction provided there is good prior biological information available [20].

Figure 3 compares the results from Bayes RC, Bayes R and GWAS analyses of bovine milk protein percentage (data described in [20]) in the region of the LALBA gene. Bayes RC incorporates prior biological information on sites that are more likely to have an effect on the trait, while Bayes R and GWAS assume all sites are equally likely to affect the trait. The top Bayes RC variant is shown in purple (with base pair position), and the strength of LD ($r^2$) between this and all other variants is colour coded (adapted from data published in [20]).

Figure 3. Comparison of results for three genome-wide analysis methods of bovine milk protein percentage in a region around the LALBA gene. (a) Bayes RC incorporates prior biological information on sites that are more likely to have an effect on the trait, while (b) Bayes R and (c) GWAS assume all sites are equally likely to affect the trait. The top Bayes RC variant is shown in purple (with base pair position), and the strength of LD ($r^2$) between this and all other variants is colour coded (adapted from data published in [20]).
A polymorphism that affects one trait is also likely to affect other traits. Thus, a multi-trait analysis may increase power to find the causal sites. Additional traits are especially useful if the causal polymorphism has a large effect on the trait because this increases statistical power. Gene expression is a trait with large effects, especially mutations acting in cis to change the expression of the allele on the same chromosome as the mutation. Another benefit of cis eQTL is that they define the gene through whose expression the QTL has its effect, thus increasing our understanding of the pathway from gene to phenotype. Other genetically simpler traits with large effects might include individual proteins and product components. An example follows. Kemper et al. [25] found that a QTL that explained 0.001 of the genetic variance for milk yield co-segregated with a QTL that explained 0.1 of the genetic variance for phosphorus concentration in milk. The SNPs with a large effect on both traits mapped near a gene for a phosphorus anti-porter that transports glucose-6-phosphate in one direction and phosphorus in the other direction across cell membranes. Glucose is a substrate for lactose synthesis, which is the major osmolarity regulator in milk and hence drives milk volume. Thus, the allele that increases milk phosphorus concentration decreases milk volume. Gene expression data showed that the same sequence variant that increases phosphorus content in milk also increases expression of this gene.

4. Genetic architecture of complex traits

Traditional SNP regression analysis of GWAS reports the estimated effect of SNPs that are declared significant. However, this information does not give a good description of the genetic architecture of the trait. In SNP regression analysis, very stringent p-values (p < 5 × 10^{-8}) are used to protect against testing as many as 1 000 000 SNPs for an effect. This has several consequences. The estimated effect of SNP's declared significant is grossly overestimated. A better estimate of the effect size can be obtained by estimating the effect of significant SNPs in an independent dataset. When this is done it is found that the significant SNPs explain a small proportion of the genetic variance estimated from pedigree or family relationships. This was called the 'missing heritability' paradox. Yang et al. [11] showed that if the combined effect of all SNPs was estimated, the proportion of genetic variance explained was much higher. For instance, the genetic variance for height in humans explained by significant SNPs was 5% of phenotypic variance, whereas all the SNPs together explained 45% of the phenotypic variance [11]. The explanation for this difference is simply that most SNP effects on height are too small to be significant given the stringent p-value used.

Even 45% is less than the 70% or more of the phenotypic variance estimated to be additive genetic variance by family studies, possibly because causal variants with low minor allele frequency (MAF) are not in high LD with any of the SNPs used. Yang et al. [26] found that imputed genome sequence explained 50% of the phenotypic variance for height. Sequence variants with MAF < 0.1 explained more variance than other MAF classes despite the fact that they were not accurately imputed. Accounting for this poor imputation of rare variants suggests that 60% of phenotypic variance is explained by genome sequence. It is also possible that the family studies overestimate the heritability due to confounding genetic effects with common family environment, or confounding additive and non-additive genetic variation. Thus, the genetic variance explained by sequence variants is almost equal to that estimated from family studies.

Bayes R provides an estimate of the number of causal variants affecting a trait and the distribution of their effects by approximating the distribution of effect sizes with a mixture of normal distributions. (We do not imply the SNP effects are literally drawn from a mixture of normal distributions, merely that this mixture can approximate almost any distribution that might describe the distribution of effect sizes.) For many traits in both humans and cattle, we find that there are thousands of SNPs with effect sizes drawn from a distribution with variance of 0.0001 σ^2 and a handful with variance 0.01 σ^2 [12–14]. Thus, complex traits are more complex than was thought, with thousands of polymorphisms, each with very small effects, affecting each trait. Their allele frequencies are biased towards low MAF compared with a neutral model, but only slightly. That is, most of the variance is due to common variants [26].

By allocating polymorphic sites into categories in our Bayes RC analysis, we can estimate the distribution of effect sizes for each category. Table 1 shows the results for milk yield in dairy cattle where SNPs were allocated to one of three categories (data described in [20]). Non-synonymous coding sites in a set of candidate genes affecting milk production had a higher proportion of non-zero effects than non-coding sites outside candidate gene regions. However, because non-coding sites are more numerous they explained most of the variance (table 1).

5. Discussion

Although the nonlinear models that we have advocated (such as Bayes R) give good predictions, they estimate or predict more variables (p) than there are subjects (n) and so there is justified concern that many other models could fit the data equally well. In Bayes R, we fix the variance for each component of the mixture to minimize the number of parameters to be estimated. There is no special reason for our choice of four components or variances of 0, 0.0001, 0.001 and 0.01 σ^2. However, we have found that a mixture of these four components can approximate a wide range of distributions. A limitation might be the absence of a normal with even smaller variance in the mixture. Nevertheless, Moser et al. [12] found that the four-component mixture could still perform well when the data were simulated under a different model. The mixing proportions themselves are assumed to be drawn from a Dirichlet distribution with the prior being equivalent to one SNP in each component of the mixture. The estimated mixing proportions are very far from the prior because the number of SNPs in the distribution with zero or small variance is much larger than the number in the component of the mixture with larger variance. As the mixture requires only four parameters, it is not surprising that the data have some power to estimate these parameters. Nevertheless, we caution against too literal an interpretation of the estimates. This caution is partially due to inherent difficulties in estimating so many variables and partially due to the use of Markov chain Monte Carlo (MCMC) methods. It is difficult to know when an MCMC chain has converged. We typically run five
Table 1. Proportion of SNP effects discovered in biologically defined SNP categories compared with the total variance explained, using a Bayes RC analysis of dairy cattle milk production. Prior to the analysis, SNPs were divided into categories based on prior knowledge of candidate genes and annotation of non-synonymous coding SNP. (Adapted from data published in [20].)

<table>
<thead>
<tr>
<th>SNP category</th>
<th>no. SNP per category (% of total)</th>
<th>proportion SNP effects per distributiona</th>
<th>total variance explained per SNP category (%)</th>
<th>( \sigma^2_{\text{g}} ) (%)</th>
<th>( \sigma^2_{\text{g}} ) (%)</th>
<th>( \sigma^2_{\text{g}} ) (%)</th>
<th>( \sigma^2_{\text{g}} ) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-synonymous coding SNP in candidate genes</td>
<td>3,726 (0.4%)</td>
<td>0.0001 ( \sigma^2_{\text{g}} ) (%)</td>
<td>55.6</td>
<td>9.9</td>
<td>0.03</td>
<td>0.0004</td>
<td></td>
</tr>
<tr>
<td>other SNP in or within 50 Kbs of candidate genes</td>
<td>57,722 (6%)</td>
<td>0.0001 ( \sigma^2_{\text{g}} ) (%)</td>
<td>99.0</td>
<td>0.4</td>
<td>0.01</td>
<td>6.7</td>
<td></td>
</tr>
<tr>
<td>all other SNP</td>
<td>947,905 (93%)</td>
<td>0.0001 ( \sigma^2_{\text{g}} ) (%)</td>
<td>99.6</td>
<td>0.4</td>
<td>0.01</td>
<td>75.7</td>
<td></td>
</tr>
</tbody>
</table>

\( \sigma^2_{\text{g}} \) is the additive genetic variance for the trait. Bayes RC estimates SNP effects as a mixture of four normal distributions: one with zero mean and variance, and the others as \( \sigma^2_{\text{g}} \), \( \sigma^2_{\text{g}} \), where \( \sigma^2_{\text{g}} \) is the additive genetic variance for the trait.

aBayes RC estimates SNP effects as a mixture of four normal distributions: one with zero mean and variance, and the others as \( \sigma^2_{\text{g}} \), \( \sigma^2_{\text{g}} \), and \( \sigma^2_{\text{g}} \), where \( \sigma^2_{\text{g}} \) is the additive genetic variance for the trait.

bThe candidate genes were a group of 790 genes selected based on differential expression in mammary gland, in experiments designed to manipulate milk production.

d cis eQTL tend to be large on average than neutral SNPs. However, a few mutations of large effect are likely to be at low MAF, most of the variance is due to QTL that individually explain a small proportion of the variance (e.g. less than 1%). Although a few mutations of large effect are likely to be at low MAF, most of the variance is due to QTL that have only slightly lower MAF on average than neutral SNPs.

The availability of dense SNP panels and genome sequence on large numbers of individuals, who have also been recorded for a complex trait, has changed our understanding of the genetics of complex traits and led to great practical benefits in the genetic improvement of livestock and crops. It now appears that thousands of polymorphisms affect a typical complex trait. The effect of these QTL varies from large to very small, but most of the variance is due to QTL that individually explain a small proportion of the variance (e.g. less than 1%). Although a few mutations of large effect are likely to be at low MAF, most of the variance is due to QTL that have only slightly lower MAF on average than neutral SNPs.

This new understanding of the genetic architecture of complex traits has implications for prediction and for identification of causal polymorphisms. The results emphasize the need for large sample sizes for both purposes. Even with access to full genome sequence, most polymorphisms explain only a tiny fraction of the variance, and therefore large sample sizes are needed to reach the conventional significance level \( p < 5 \times 10^{-8} \).

The strategy for prediction depends on the \( N_e \) of the population. For populations with small recent \( N_e \) (livestock, some crops), a prediction equation derived by BLUP based on a moderately dense SNP panel works well. However, even here this approach has disadvantages. The prediction equation is generally not robust to minor changes in the population such as a change in breed, place or time. For instance, a prediction equation trained on one breed of livestock has little accuracy in another, closely related breed [6]. An alternative strategy is to use a nonlinear method and a multi-breed training population (or more genetically diverse population) in the hope that this will lead to more robust predictions.

For populations with large recent \( N_e \), BLUP prediction of genetic value requires enormous training population size. As a population of sufficient size is often not available, some attempt is usually made to reduce the amount of the genome that is considered (e.g. by using coding variants only). However, a better alternative would be to aim for good coverage of the genome and to use nonlinear prediction methodology and the strategies discussed above to maximize the accuracy.

If the SNP panel does not explain all of the genetic variance, there is a limit to the accuracy that can be achieved. To overcome this limit, genome sequence data, which should contain the causal variants, can be used. However, to extract extra accuracy from genome sequence data requires a nonlinear statistical method [27,28].

To estimate the effect of all sequence variants simultaneously is a challenging task. We have illustrated above two strategies that can be used to help with this task. First, a QTL with a small effect on one trait may have a larger effect on another trait. Therefore, a multi-trait analysis increases power to find sequence variants which have an effect on any trait. Gene expression is a particularly useful type of trait because the effects of cis eQTL tend to be large
and the result immediately indicates the gene through which the polymorphism acts.

Second, external information on sites in the genome that, if mutated, would have an effect on phenotype, such as the ENCODE data [29], are useful in deciding which sites are most likely to be causal. Variants that change the amino acid sequence of proteins are more likely to affect phenotype than random sites in the genome, and this is used in the Bayes RC method described above. However, evidence is mounting that the majority of mutations that give rise to variation in complex traits reside in regulatory elements that alter gene expression [30–32] (reviewed by Pai et al. [33]).

Cis-acting elements affect gene expression only on the same DNA molecule, thus acting in an allele-specific manner. Detecting allele-specific expression (ASE) is an alternative method of finding cis eQTL [34]. ASE occurs at heterozygous variants where one allele is more highly expressed in the mRNA than the other. Recently, Crowley et al. [35] and Chamberlain et al. [23] reported that 89% of mice and cattle genes, respectively, show ASE in at least one tissue. Therefore, it seems likely that cis eQTL are very common.

Sites in the genome that would affect gene expression if mutated can also be identified by genomic features such as histone marks and transcription factor binding sites, which are indicative of enhancers [36], transcription start sites [36] or promoters [37], any of which could be involved in gene regulation, and which therefore might affect phenotype if mutated. A consortium called Functional Annotation of ANimal Genomes (FAANG) [38] is now planning to annotate livestock genomes for such histone marks as well as other markers of open chromatin in a similar fashion to that in humans (the ENCODE consortium).

Authors’ contributions. All authors contributed to the writing of this review and to the original research, which has been reported elsewhere.

Competing interests. We declare we have no competing interests.

Funding. This work was supported in part by Australian Research Council grant DP130100563.

References

15. Schaeffer LR. 2006 Strategy for applying genome-wide association mapping that accounts for multiple association mapping that accounts for multiple


