agriculture

Review

Identification of Genomic Variants Causing Variation in Quantitative Traits: A Review

Theo Meuwissen [1], Ben Hayes [2], Iona MacLeod [3] and Michael Goddard [3,4,*]

[1] Faculty of Biosciences, Norwegian University of Life Sciences, P.O. Box 5003, 1432 As, Norway
[2] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia, QLD 4072, Australia
[3] Agriculture Victoria Research, Agribio, Bundoora, VIC 3083, Australia
[4] School of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, VIC 3011, Australia
* Correspondence: mike.goddard@agriculture.vic.gov.au

Abstract: Many of the important traits of livestock are complex or quantitative traits controlled by thousands of variants in the DNA sequence of individual animals and environmental factors. Identification of these causal variants would be advantageous for genomic prediction, to understand the physiology and evolution of important traits and for genome editing. However, it is difficult to identify these causal variants because their effects are small and they are in linkage disequilibrium with other DNA variants. Nevertheless, it should be possible to identify probable causal variants for complex traits just as we do for simple traits provided we compensate for the small effect size with larger sample size. In this review we consider eight types of evidence needed to identify causal variants. Large and diverse samples of animals, accurate genotypes, multiple phenotypes, annotation of genomic sites, comparisons across species, comparisons across the genome, the physiological role of candidate genes and experimental mutation of the candidate genomic site.

Keywords: genomic prediction; causal variants; linkage disequilibrium; quantitative trait loci

check for updates

Citation: Meuwissen, T.; Hayes, B.; MacLeod, I.; Goddard, M. Identification of Genomic Variants Causing Variation in Quantitative Traits: A Review. *Agriculture* **2022**, *12*, 1713. https://doi.org/10.3390/agriculture12101713

Academic Editor: Ligang Wang

Received: 5 September 2022
Accepted: 10 October 2022
Published: 17 October 2022

1. Introduction

Most of the traits that are important in livestock and crops are quantitative or complex traits. Great improvement in these traits has been accomplished by selecting animals or plants based on their phenotype and that of their relatives. In the last decade the rate of genetic improvement has been increased by genomic selection or genomic prediction (GP) [1]. The purpose of this review is to consider the value of knowledge about casual variants (CVs) in genomic selection for complex traits in livestock. Three key aspects are considered: is it worthwhile, how might we identify them and the success to date.

2. What Are the Advantages of Identifying Causal Variants?

We consider four possible benefits: more accurate GP, knowledge of the physiology of the trait, knowledge of the evolution of the genomic sites controlling the trait, and to provide targets for gene-editing.

More accurate genomic prediction. Genomic prediction (GP) is the prediction of breeding value from genotypes at genetic markers, such as single nucleotide polymorphisms (SNP) scattered throughout the genome. A training population recorded for the trait and genotyped for the markers is used to estimate a prediction equation that takes marker genotypes as input and outputs estimated breeding values. This prediction equation can then be used to improve the prediction of breeding value in selection candidates.

The prediction equation is typically linear in the marker genotypes, like a multiple regression equation, and it is tempting to interpret the regression coefficient of a marker as the effect of that marker on the trait. However, this is incorrect. The markers usually do not cause variation in the trait but are in linkage disequilibrium (LD) with the genetic variants that do cause variation in the trait. GP works because the markers are sufficiently dense so that the genotypes at most casual variants can be predicted from the genotypes at the markers.

The accuracy of GP might be improved by using causal variants in 3 ways: capturing more of the genetic variance, estimating marker effects more accurately and avoiding the reduction in accuracy due to recombination causing changes in LD pattern.

A panel of markers may not perfectly predict the genotypes at all causal variants. For instance, causal variants that have a low allele frequency cannot be in high LD with markers that have a high minor allele frequency [2]. Consequently the markers do not track all the genetic variance and do not use all of it in their prediction. In human genetics the genetic variance explained by SNPs is typically 1/3 to 2/3 of the genetic variance estimated from pedigree analysis e.g., [3,4]. In livestock, when the model includes a genetic effect following the markers and one following the pedigree relationship, the latter explains 1–50% of the total genetic variance [5–7]. The proportion explained by the markers varies with the diversity of the population: if it includes multiple breeds then denser markers are needed to capture all the causal variants. This model, with both an effect explained by markers and one following pedigree, is also used in large scale estimation of breeding values and again with 10–50% of the variance assumed to follow the pedigree relationship. If the markers explain a fraction r^2 of the genetic variance, the maximum accuracy of a prediction based on these markers is r.

This maximum accuracy is not achieved in practice because the marker effects are not estimated with perfect accuracy. The accuracy depends on the 'effective number of chromosome segments' (Me) segregating in the population [8–10]. This number is low in most livestock breeds because they have a low recent effective population size ([11–13]. For instance, in Holsteins Me has been estimated at around 3000 to 7000 [9,10]). If the number of causal variants was much less than the number of effective chromosomal segments, we expect that their effects could be estimated more accurately than those of the markers leading to more accurate GP Estimated Breeding Values (EBVs) [14]. Evidence suggests the number of causal variants is >4000 for most traits, so the accuracy of their estimated effects might be slightly greater than that of markers in a single breed analysis. However, in a more diverse population, such as a mixture of breeds, the number of effective chromosomal segments is larger and so the advantage of using causal variants might be higher [14].

The accuracy of GP is eroded if the LD in the target population is different to that in the training population. For instance, LD changes over time due to recombination and it differs between parts of a population if the population is not panmictic. Consequently, prediction accuracy is not robust over time and space.

These predictions of accuracy using causal variants are largely borne out by simulation studies [14,15]. However, simulation studies may not simulate the real world. We cannot test the advantage of using causal variants in real data because we do not know what they are. The best we can do at present is to test methods that attempt to find large sets of markers closer to the causal variants than those on the panels normally used [16–18].

In real data, several studies have demonstrated an increase in the accuracy of genomic prediction through use of selected sequence variants that were identified as being close to causal variants e.g., [19–24]. These studies generally found the advantage of adding sequence variants to marker panels was most apparent for mixed breed reference populations and/or for prediction into different breeds or crossbreds. That is, the predictions held their accuracy better in animals less related to the training populations, compared to using markers from a standard panel.

This indicates that there is indeed an advantage in the real world for attempting to identify causal variants, or markers closer to CV, to improve robustness of genomic prediction for individuals less closely related to training populations. The major challenge is the large number of causal variants that need to be identified across the wide range of economically important complex traits.

Knowledge of the physiology, gene editing and evolution of the trait. Knowing how a change in DNA sequence affects a complex trait such as milk yield is of great scientific interest. The first step might be to identify the gene through which the mutation acts. If the markers are dense enough it may be possible to guess the gene from a marker that is associated with the trait. However, knowledge of the causal variant would help discover the gene and the way in which the causal variant affects it (e.g., by changing the protein sequence or regulation of expression). In human medicine, knowledge of the gene without the causal variant may be enough to suggest a drug target to treat a disease. This could also be the case in livestock diseases. Knowledge of the gene and the causal variant may be used as targets for gene-editing [25].

The large-scale use of gene-editing for genetic improvement requires a large panel of target sites with causal effects [26]. For gene-editing purposes, it may however suffice to know the gene and whether it needs to be up-regulated or down-regulated, or whether it's functionality needs to be reduced or enhanced. I.e., it may not require knowing the causal variant, although this would be of great help. Also, narrowing the causal variants down to a set of approximately 10 potentially causative variants would be helpful for gene-editing, where all 10 variants could be edited and tested for their effects.

It is also of scientific interest to know how the mutations affecting complex traits evolve. For instance, does domestication lead to fixation of mutants that would be deleterious in the wild or does it involve a change in allele frequencies at many loci? This cannot be studied unless we know the causal variant because a marker in LD with the causal variant may not share the same evolutionary history.

3. Why Is It Hard to Identify Causal Variants?

Success to date for unequivocally identifying causal mutations for complex traits in livestock is limited and has generally been restricted to variants that have relatively large effects e.g., [27–30]. However, large databases of livestock sequences (e.g., 1000 Bull Genomes Project and SheepGenomesDB [31,32]) have enabled imputation to sequence of many thousands of animals with phenotypes. As a result, there are a growing number of published studies that have used imputed whole-genome sequence to identify putative causal variants.

However, it is still difficult to identify causal variants because they are in LD with other DNA variants and their effects are usually small. If two variants are in complete LD it is impossible to tell from genetic data which is responsible for an effect on a trait. Even if the LD is not complete, if the effect is small, enormous sample size is needed to be confident which is causal and which is associated with the trait due to LD.

Other evidence, discussed below, such as that a mutation alters the activity of a protein, may help build a case that it is causal. However, the 'gold standard' proof that a mutation affects the trait, is to make a transgenic individual and show that the phenotype is recapitulated. This is seldom practical in livestock and never in humans, although gene-edited tissue-cultures may reveal some evidence.

4. Evidence for Causality

Since it is usually not possible to achieve the gold standard proof that a variant is causal, we attempt to mount enough evidence that a variant is most likely causal [33].

A common starting point to identify causal variants for complex traits is to undertake a genome wide association study. The data that is collected for the training population in GP is the same as the data used in a genome wide association study (GWAS) to map the causal variants to a part of the genome. However, high precision mapping requires higher marker-density than GP, preferably sequence genotypes, and this may be achieved by imputation of the missing high-density genotypes. Bayesian GP methods that allow some markers to have no effect on the trait can also be used to map causal variants and to describe the genetic architecture of the trait [34–36]. If all sequence variants are included in the data, then the analysis can potentially identify the causal variants. An output of such analyses is the probability that a variant is included in the model. If all sequence variants are available in the data then a sequence variant that has a probability of 100% to affect the trait is supposedly causal. In human genetics this process is called fine scale mapping and is usually applied to a segment of the genome in which it is believed a causal variant exists. This seldom leads to a single causal variant but more likely to a set of variants which is believed with 90% probability to include the causal variant. Even this conclusion may be wrong if the true causal variant is not included in the analysis. This is likely for classes of causal variants that are difficult to genotype and impute such as structural variants.

To build the case that a variant is causal there are 8 types of data which are helpful—larger sample size, use of actual instead of imputed genotypes, other traits which map to the same location, annotation of genomic sites, comparisons across species, comparisons between parts of the genome, genes with a known role in the physiology of the trait and experimental mutation of the site.

4.1. Increase Sample Size and Diversity

Obviously increasing sample size increases power to distinguish between variants that are not in complete LD. Increasing the genetic diversity of the sample (e.g., by using multiple breeds) decreases the LD and so increases the probability of distinguishing between sites in LD and causal effects. An approach to increase sample size, now gaining popularity in livestock, is a meta-GWAS that combines the summary statistics from a number of individual GWAS studies e.g., [37–39]. The major advantage of this approach in addition to increasing power and diversity, is that it alleviates the difficulties associated with sharing raw data across groups and countries.

Despite large sample size, a SNP other than the CV may be more significant than the CV due to sampling error. Consider a region with a single CV and compare the CV with a SNP that is in LD with the CV. What is the probability that the SNP is more highly significant than the CV? Let b_{CV} = the estimated effect of the CV and b_{SNP} = the estimated effect of the SNP. Then $b_{CV} - b_{SNP}$ ~N(b(1 − r),(1 − r) s^2/(Npq)) where b = true effect of the CV, r = the LD i.e., the correlation between the CV and the SNP, s = standard deviation of the residuals, N = sample size, p and q = 1 − p are the allele frequencies at both the CV and the SNP, which are assumed to be the same. Therefore, the probability that $b_{SNP} > b_{CV}$ is the probability that x~N(0,1) > tsqrt(1 − r) where t = bsqrt(Npq)/s is a t statistic for the true effect of the CV.

Table 1 shows how this probability varies with the LD between the SNP and CV (r) and the true t-value for the CV (t). For instance, if a CV explains 0.0001 of the phenotypic variance and we have a sample size N = 100,000 then the E(t) = sqrt(10). (If the CV explains 0.01 of the phenotypic variance but N = 1000, then E(t) is also sqrt(10)). From the table if t = 3 and r = 0.94, the probability that the SNP is more significant than the CV is 0.23. This probability is the probability that a single SNP is more significant than the CV. If this probability = P then the probability that one of n conditionally independent SNPs (conditional on their correlation to the CV) is more significant than the CV = $1 - (1 - P)^n$. This probability is high if r is close to 1 and n is high. The number of conditionally independent SNPs may be seen as an effective number of SNPs that are in high LD with the CV, which may be smaller than the actual number of SNPs that are in high LD with the CV, especially when these SNPs are incorporated in LD blocks.

In a Bayesian analysis the choice of variant as the putative CV depends on the posterior probability which in turn depends on the likelihood and the prior probability (π). The difference in log(likelihood) between a CV and a SNP in LD with it is:

$$\log(\pi_{CV}/\pi_{SNP}) + 0.5 \times t^2 \times (1 - r)^2 \quad (1)$$

Thus if t is small and r approaches 1, the choice of the variant as the putative CV depends on the priors. The use of prior information that identifies potential CVs (see section ‘Annotation of genomic sites’) may thus be important in Bayesian analyses.

After a Bayesian analysis is conducted and reveals a quantitative trait locus (QTL) region, we can calculate the difference in posterior probability (PP) between the putative CV, i.e., the highest PP in the QTL region and the second highest PP. This reveals the (log) odds ratio of the putative CV being the true CV versus the second highest PP pointing to the CV. Also, it is possible to identify a set of SNPs that collectively give a PP > 0.9 as a 90% confidence set that is likely to contain the CV.

Table 1. The probability that a SNP in LD with the CV is more significant than the CV. (t = true t-value for the CV, r = LD correlation between CV and SNP).

t	r	0.5	0.75	0.875	0.9375	0.96875	0.984375	0.992188	0.996094
1		0.23975	0.308538	0.361837	0.401294	0.429842	0.450262	0.464784	0.475082
2		0.07865	0.158655	0.23975	0.308538	0.361837	0.401294	0.429842	0.450262
3		0.016947	0.066807	0.144422	0.226627	0.297942	0.35383	0.395441	0.425634
4		0.002339	0.02275	0.07865	0.158655	0.23975	0.308538	0.361837	0.401294
5		0.000203	0.00621	0.03855	0.10565	0.18838	0.265986	0.329266	0.37733
6		1.1×10^{-5}	0.00135	0.016947	0.066807	0.144422	0.226627	0.297942	0.35383
7		3.72×10^{-7}	0.000233	0.006664	0.040059	0.107962	0.190787	0.268051	0.330874
8		7.71×10^{-9}	3.17×10^{-5}	0.002339	0.02275	0.07865	0.158655	0.23975	0.308538

4.2. Use of Actual Instead of Imputed Genotypes

Imputed genotypes may show reduced trait-associations due to imputation inaccuracies. The latter implies that a causal variant, whose genotypes are imputed, may show a lower GWAS signal than another site that is merely in LD with the causal site [40]. It is thus suggested to use accurate, actual genotypes instead of imputed genotypes when trying to distinguish between causal and LD sites.

4.3. Multiple Trait Analysis

If a variant affects multiple traits then multi-trait analysis increases power in a similar way to increasing sample size (For different approaches to multi-trait analysis see [41,42]). This is particularly useful if the causal variant has a large effect on one of the traits. For instance, [43] found that a small effect on milk yield was associated with a large effect on milk phosphorus concentration and this locus had a large effect on the expression of the gene SLC37A1.

One class of traits which may be useful is the expression of genes that can be measured using RNA sequencing. Variants affecting gene expression are called expression QTL (eQTL). An allele of an eQTL may affect the expression of a gene on the same chromosome (cis eQTL) or the expression of the gene from both homologous chromosomes (trans eQTL). cis eQTL are located close (usually <1 mb) to the gene they regulate and typically have large effects on the expression of the gene. cis eQTL can be mapped with a smaller sample size than most QTL because they have large effects. However, moderate sample sizes are still needed. For instance, 1000 individuals measured for a cis eQTL that explains 10% of phenotypic variance in expression of the gene, gives the same power as 100,000 individuals for a QTL explaining 0.1% of the phenotypic variance in a quantitative trait. Considerations for eQTL studies are which tissue and timepoint on which to measure gene expression. Trans eQTL, which affect the expression on genes on other chromosomes typically have smaller effects than cis eQTL.

4.4. Annotation of Genomic Sites

Genomic sites are annotated in several ways and this can be helpful in evaluating the likelihood of causality. For instance, sites can be coding or non-coding and within coding, they may be synonymous or non-synonymous. We assume that non-synonymous coding sites are more likely to affect a trait than other sites but this may not be correct. Two of the best-known variants affecting milk production in dairy cattle are thought to be coding variants in DGAT1 and GHR [44,45].

While there is considerable annotation now available for genic regions in livestock, there is still relatively little known about the function of intergenic sites. In human genetics, the ENCODE and Roadmap projects have provided publicly available resources listing functional regions in the human genome [46,47]. The Functional Annotation of Animal Genomes (FAANG) global collaboration aims to provide a similar resource for livestock [48] Many of these annotations are based on assays that identify parts of the genome with a function such as open chromatin, histone marks, transcription factor binding sites. Using a small number of individuals and often multiple tissues, these types of annotation identify very localised regions genome-wide that have an influence on gene expression. The annotations can be specific to tissues, developmental stages, rearing conditions, or the disease status of the animal. Although there have been some attempts to lift over such annotations from the human genome this has not generally provided high enough resolution [49].

As described here the annotation of genomic sites does not rely on genetic variation in them and so does not suffer from LD in the way that analysis of genetic differences in a trait does. Also, it means that it is only necessary to assay a small number of animals. However, there are a great many of these sites in the genome and it is not clear which if any of them would affect a particular trait of interest. Neither is it obvious how genetic variants within the region might affect their function. For instance, Chipseq assays for methylation 'tags' on histones are thought to identify genome regions of 200–1000 bp that are enhancers and promoters influencing gene expression. A SNP that lies within such a region might affect the function of the enhancer or promoter but it might have no effect on that function. We can compare animals with different genotypes at this SNP and determine whether or not the genotype affects the assay result. If it does affect the assay, the SNP may also affect the expression of the gene and hence economically important phenotypes. However, this requires relatively larger sample sizes and if there are multiple SNPs in LD it may still be difficult to tell which is causal. That is, this is a genetic analysis of a new trait defined by an assay for a function in the DNA. In this respect it is similar to expression QTL which are polymorphisms that affect gene expression. Ideally, we would like to combine an assay that identifies a specific region of the genome as functional with genetic evidence that a polymorphism in that region affects its function.

Although functional annotations are not trait specific, they have been shown to be enriched for putative causal variants discovered from trait specific GWAS [22,49–51]. Therefore, when considered jointly with the effect of genetic variants on specific traits, these annotations are a valuable tool towards identifying causal variants. Below we consider how to appropriately weight this information.

4.5. Comparisons across Species

If the same allele is conserved at a site across many species it must be subject to selection and therefore must have some function. Such conserved sites are enriched among sites affecting complex traits [16].

4.6. Comparisons across the Genome

If there is a phenotype that varies across the genome it is possible to learn the DNA sequence associated with the phenotype. For instance, assays can detect regions of open chromatin by their hypersensitivity to DNase (DHS regions). By comparing DNA sequences under DHS regions with those not under DHS regions you can identify sequences that lead to these sites and variations in the sequence that cause an increase or decrease in the probability of such a region [52]. This process identifies sites that affect a molecular phenotype and it does so without the confusion caused by LD. However, there is no proof that these sites affect a phenotype in which we are interested.

4.7. Genes with a Role in the Physiology

If a mutation is proposed to affect a complex trait through a given gene it adds to the evidence if that gene has a known role in the trait. This was the case for the two milk production QTL affecting the protein coding sequence of DGAT1 and GHR.

4.8. Experimental Mutation of the Site

Only rarely will we make a transgenic animal to prove that a genomic variant is causal for the trait of interest. However, we can test transformed cell lines for a molecular phenotype such as gene expression. This has been done for a single proposed mutation (e.g., DGAT1) but can now be done for thousands of sites in massively parallel reporter assays [53] The effect of a regulatory variant may be tissue specific so it may be necessary to have cell lines from multiple tissue types.

5. Combining Information from Different Sources

Given many sources of information which might predict which sites are likely to have an effect on phenotype, it is beneficial to construct a multiple regression equation to predict the probability that a site affects phenotype. The method called Bayes RC is a Bayesian method in which genetic markers can be classified according to the annotations they have [35]. Then the probability that each class of markers is associated with the trait is estimated. Potentially a multiple regression equation could be used instead of a classification.

A common method in human genetics is stratified LD score regression [54]. This uses the chi-square statistic for each marker in a single SNP regression analysis of GWAS data. This measures the variance of the trait associated with the marker which may also indicate the proportion of similar markers that have a non-zero effect on the trait. In a single SNP regression GWAS, the apparent effect of the SNP is due to the SNP itself and all those in LD with it. Therefore, in LD score regression the independent variable is the sum of LD r^2 between the focal SNP and all surrounding SNPs. In stratified LD score regression separate LD scores are calculated for each annotation of the surrounding SNPs.

Another method is to define different genomic relationship matrices among all the individuals for each category of genetic markers [16,55]. For instance, a genomic relationship matrix (GRM) based on coding SNPs and one based on random SNPs. Then it is possible to estimate the genetic variance associated with each type of GRM thus indicating which annotations identify markers causing the most variance in a complex trait.

Xiang et al. [16] illustrate some of these approaches. They developed a score (called FAETH) for polymorphic sites in cattle based on a number of annotations and combined this with multi-trait genetic analysis to find approximately 50,000 SNPs that were more likely to be causal or close to causal variants. A SNP chip containing these SNPs gave higher accuracy of genomic EBVs than previous SNP panels.

6. Creditable Sets Instead of Single Causal Variants

The focus in this review is on complex or quantitative traits but the same problems occur in identifying the mutation causing phenotypes that can be caused by a single mutation such as many genetic abnormalities. The effect size in this case is large, so the sample size needed is smaller but the problem of LD between a causal variant and other variants is the same. Perhaps these mutations are easier to identify than those for complex traits because they are often coding mutations. However, almost never is the transgenic animal made to confirm we have identified the correct mutation. Therefore, we should be able to build an equally strong case that we have identified the causal mutation for a complex trait as we do for Mendelian traits provided we increase sample size. Despite this, success in identifying variants affecting complex traits has been low.

In some chromosomal regions, mutations at several sites may cause similar effects, e.g., due to affecting the expression of a gene, or reducing the functionality of a gene's transcript. For instance, in the DGAT1 region, next to the known site, other sites may have similar trait effects. In such cases, attempts to find 'the' causal mutation will at best result in the discovery of the biggest of the mutations, but the conclusion that herewith 'the' mutation is found is wrong.

In view of the latter, and the difficulty in finding 'the' causative mutation (if it exists), a useful aim for a GWAS study may be find a set of e.g., 10 potential causal variants for every QTL. The latter will affect our aims for the detection of causal variants:

- Accuracy of GP: all 10 variants will be in very high LD with the causative mutation, and the LD is not expected to change markedly with genetic distances, i.e., the reduction in GP accuracy of having a set of 10 potential instead of 1 causal variant will be limited. Genotyping costs will be increased, but genotyping costs are generally small.
- Knowing the gene that affects the trait without knowing the causal variant will be useful for the study of the trait physiology but not as useful as also knowing the causal variant.
- The same holds for the evolution of the sites.
- A set of 10 potential causal variants will enhance the costs of the initial stages of a gene-editing program, where the effect of the gene-edit on the trait is tested. This stage will require 10 such tests instead of 1. However, if the causal variant is not amongst the set of 10 potential causative variants, the gene-editing program will not be successful.

It seems that the accuracy of GP and the gene-editing results are little affected by having a set of 10 potential instead of 1 causal variant, as long as the actual true causal variant is amongst these 10. However, the study of the trait physiology and site evolution will be compromised.

7. The Future

We have argued above that it would be beneficial to identify the genomic variants causing variation in quantitative traits. Although success to date is limited, we believe that the opportunity exists for greater success in the near future by combining the approaches discussed above. Increasing sample size is being achieved through the commercial use of GP but this could be accelerated by international collaboration to build larger and more diverse data sets. International collaboration is already contributing to annotation of livestock genomes, for instance, through FAANG and livestock GTEx. Two further improvements are now within reach—identification of structural variants and massively parallel reporter assays (MPRA). Most current genotype data is on SNPs but it is likely that causal variants include structural variants. Using short read sequencing it has been hard to call genotypes at structural variants but the use of long read sequencing should improve this situation. MPRA test the effect of specific mutations uncomplicated by LD but they require a phenotype that can be measured in vitro such as gene expression. An approach which has been underutilized is discovering functional genome sequences by comparing parts of the genome [56]. This requires a phenotype that is associated with a specific location in the genome, for instance, the height of ChiPseq peaks. By comparing the sequence under Chipseq peaks, it is possible to discover the sites that determine where these functional elements occur. This leads to identification of causal variants without complication from LD but does not directly target phenotypes that can only be observed on whole animals.

Author Contributions: All authors contributed to this review in areas where they were expert. All authors have read and agreed to the published version of the manuscript.

Funding: TM thanks Norwegian Research Council for funding in project 309611.

Acknowledgments: We would like to thank Mehrnush Forutan for help with references.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Meuwissen, T.H.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829. [CrossRef] [PubMed]
2. Wray, N.R. Allele frequencies and the r2 measure of linkage disequilibrium: Impact on design and interpretation of association studies. *Twin. Res. Hum. Genet.* **2005**, *8*, 87–94. [CrossRef] [PubMed]
3. Yang, J.; Benyamin, B.; McEvoy, B.P.; Gordon, S.; Henders, A.K.; Nyholt, D.R. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **2010**, *42*, 565–569. [CrossRef]
4. Speed, D.; Cai, N.; Johnson, M.R.; Nejentsev, S.; Balding, D.J. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **2017**, *49*, 986–992. [CrossRef] [PubMed]
5. Zhang, C.; Kemp, R.A.; Stothard, P.; Wang, Z.; Boddicker, N.; Krivushin, K.; Dekkers, J.; Plastow, G. Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. *Genet. Sel. Evol.* **2018**, *50*, 14. [CrossRef] [PubMed]
6. Jensen, J.; Su, G.; Madsen, P. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. *BMC Genet.* **2012**, *13*, 44. [CrossRef]
7. Haile-Mariam, M.; Nieuwhof, G.J.; Beard, K.T.; Konstatinov, K.V.; Hayes, B.J. Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. *J. Anim. Breed. Genet.* **2013**, *130*, 20–31. [CrossRef] [PubMed]
8. Goddard, M. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* **2009**, *136*, 245–257. [CrossRef] [PubMed]
9. van den Berg, S.; Calus, M.P.L.; Meuwissen, T.H.E.; Wientjes, Y.C.J. Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. *BMC Genet.* **2015**, *16*, 146. [CrossRef] [PubMed]
10. Erbe, M.; Gredler, B.; Seefried, F.R.; Bapst, B.; Simianer, H. A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLoS ONE* **2013**, *8*, e81046. [CrossRef] [PubMed]
11. Uimari, P.; Tapio, M. Extent of linkage disequilibrium and effective population size in Finnish Landrace and Finnish Yorkshire pig breeds. *J. Anim. Sci.* **2011**, *89*, 609–614. [CrossRef]
12. Kijas, J.W.; Lenstra, J.A.; Hayes, B.; Boitard, S.; Porto Neto, L.R.; San Cristobal, M.; Servin, B.; McCulloch, R.; Whan, V.; Gietzen, K. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* **2012**, *10*, e1001258. [CrossRef] [PubMed]
13. Bovine HapMap, C.; Gibbs, R.A.; Taylor, J.F.; Van Tassell, C.P.; Barendse, W.; Eversole, K.A.; Gill, C.A.; Green, R.D.; Hamernik, D.L.; Kappes, S.M. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **2009**, *324*, 528–532. [CrossRef]
14. MacLeod, I.M.; Hayes, B.J.; Goddard, M.E. The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data. *Genetics* **2014**, *198*, 1671–1684. [CrossRef] [PubMed]
15. Meuwissen, T.; Hayes, B.; Goddard, M. Genomic selection: A paradigm shift in animal breeding. *Anim. Front.* **2016**, *6*, 6–14. [CrossRef]
16. Xiang, R.; Van Den Berg, I.; MacLeod, I.M.; Hayes, B.J.; Prowse-Wilkins, C.P.; Wang, M.; Bolormaa, S.; Liu, Z.; Rochfort, S.J.; Reich, C.M. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 19398–19408. [CrossRef]
17. Kichaev, G.; Yang, W.-Y.; Lindstrom, S.; Hormozdiari, F.; Eskin, E.; Price, A.L.; Kraft, P.; Pasaniuc, B. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genet.* **2014**, *10*, e1004722. [CrossRef] [PubMed]
18. Kircher, M.; Witten, D.M.; Jain, P.; O'Roak, B.J.; Cooper, G.M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **2014**, *46*, 310–315. [CrossRef]
19. Brøndum, R.F.; Su, G.; Janss, L.; Sahana, G.; Guldbrandtsen, B.; Boichard, D.; Lund, M.S. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *J. Dairy Sci.* **2015**, *98*, 4107–4116. [CrossRef] [PubMed]
20. Moghaddar, N.; Khansefid, M.; van der Werf, J.H.J.; Bolormaa, S.; Duijvesteijn, N.; Clark, S.A.; Swan, A.A.; Daetwyler, H.D.; MacLeod, I.M. Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. *Genet. Sel. Evol.* **2019**, *51*, 72. [CrossRef]
21. Khansefid, M.; Goddard, M.E.; Haile-Mariam, M.; Konstantinov, K.V.; Schrooten, C.; de Jong, G.; Jewell, E.G.; O'Connor, E.; Pryce, J.E.; Daetwyler, H.D. Improving Genomic Prediction of Crossbred and Purebred Dairy Cattle. *Front. Genet.* **2020**, *11*, 598580. [CrossRef] [PubMed]
22. Xiang, R.; MacLeod, I.M.; Daetwyler, H.D.; de Jong, G.; O'Connor, E.; Schrooten, C.; Chamberlain, A.J.; Goddard, M.E. Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. *Nat. Commun.* **2021**, *12*, 860. [CrossRef]
23. VanRaden, P.M.; Tooker, M.E.; O'Connell, J.R.; Cole, J.B.; Bickhart, D.M. Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet. Sel. Evol.* **2017**, *49*, 32. [CrossRef]
24. Al Kalaldeh, M.; Gibson, J.; Duijvesteijn, N.; Daetwyler, H.D.; MacLeod, I.; Moghaddar, N.; Lee, S.H.; Van Der Werf, J.H. Using imputed whole-genome sequence data to improve the accuracy of genomic prediction for parasite resistance in Australian sheep. *Genet. Sel. Evol.* **2019**, *51*, 1–13. [CrossRef] [PubMed]

25. Bishop, T.F.; Van Eenennaam, A.L. Genome editing approaches to augment livestock breeding programs. *J. Exp. Biol.* **2020**, *223*, jeb207159. [CrossRef]
26. Jenko, J.; Gorjanc, G.; Cleveland, M.A.; Varshney, R.K.; Whitelaw, C.B.A.; Woolliams, J.A.; Hickey, J.M. Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. *Genet. Sel. Evol.* **2015**, *47*, 55. [CrossRef] [PubMed]
27. Johnsson, M.; Jungnickel, M.K. Evidence for and localization of proposed causative variants in cattle and pig genomes. *Genet. Sel. Evol.* **2021**, *53*, 67. [CrossRef] [PubMed]
28. Tellam, R.; Cockett, N.; Vuocolo, T.; Bidwell, C. Genes Contributing to Genetic Variation of Muscling in Sheep. *Front. Genet.* **2012**, *3*, 164. [CrossRef]
29. Kambadur, R.; Sharma, M.; Smith, T.P.L.; Bass, J.J. Mutations in myostatin (GDF8) in Double-Muscled Belgian Blue and Piedmontese Cattle. *Genome Res.* **1997**, *7*, 910–915. [CrossRef] [PubMed]
30. McPherron, A.C.; Lee, S.-J. Double muscling in cattle due to mutations in the myostatin gene. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 12457–12461. [CrossRef] [PubMed]
31. Hayes, B.J.; Daetwyler, H.D. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annu. Rev. Anim. Biosci.* **2019**, *7*, 89–102. [CrossRef]
32. Daetwyler, H.D.; Brauning, R.; Chamberlain, A.J.; McWilliam, S.; McCulloch, A.; Vander Jagt, C.J.; Bolormaa, S.; Hayes, B.J.; Kijas, J.W. 1000 Bull Genomes and SheepGenomesDB projects: Enabling cost-effective sequence level analyses globally. In Proceedings of the 22nd Australian Association for Animal Breeding and Genetics, Townsville, Australia, 2–5 July 2017.
33. Ron, M.; Weller, J.I. From QTL to QTN identification in livestock—Winning by points rather than knock-out: A review. *Anim. Genet.* **2007**, *38*, 429–439. [CrossRef]
34. Kemper, K.E.; Reich, C.M.; Bowman, P.J.; vander Jagt, C.J.; Chamberlain, A.J.; Mason, B.A.; Hayes, B.J.; Goddard, M.E. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genet. Sel. Evol.* **2015**, *47*, 29. [CrossRef]
35. MacLeod, I.M.; Bowman, P.J.; Vander Jagt, C.J.; Haile-Mariam, M.; Kemper, K.E.; Chamberlain, A.J.; Schrooten, C.; Hayes, B.J.; Goddard, M.E. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genom.* **2016**, *17*, 144. [CrossRef]
36. Moser, G.; Lee, S.H.; Hayes, B.J.; Goddard, M.E.; Wray, N.R.; Visscher, P.M. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLoS Genet.* **2015**, *11*, e1004969. [CrossRef]
37. Pausch, H.; Emmerling, R.; Gredler-Grandl, B.; Fries, R.; Daetwyler, H.D.; Goddard, M.E. Meta-analysis of sequence-based association studies across three cattle breeds reveals 25 QTL for fat and protein percentages in milk at nucleotide resolution. *BMC Genom.* **2017**, *18*, 853. [CrossRef]
38. van den Berg, I.; Xiang, R.; Jenko, J.; Pausch, H.; Boussaha, M.; Schrooten, C.; Tribout, T.; Gjuvsland, A.B.; Boichard, D.; Nordbø, Ø.; et al. Meta-analysis for milk fat and protein percentage using imputed sequence variant genotypes in 94,321 cattle from eight cattle breeds. *Genet. Sel. Evol.* **2020**, *52*, 37. [CrossRef]
39. Bouwman, A.C.; Daetwyler, H.D.; Chamberlain, A.J.; Ponce, C.H.; Sargolzaei, M.; Schenkel, F.S.; Sahana, G.; Govignon-Gion, A.; Boitard, S.; Dolezal, M.; et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat. Genet.* **2018**, *50*, 362–367. [CrossRef]
40. Pausch, H.; MacLeod, I.M.; Fries, R.; Emmerling, R.; Bowman, P.J.; Daetwyler, H.D.; Goddard, M.E. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genet. Sel. Evol.* **2017**, *49*, 24. [CrossRef]
41. Bolormaa, S.; Pryce, J.E.; Reverter, A.; Zhang, Y.; Barendse, W.; Kemper, K.; Tier, B.; Savin, K.; Hayes, B.J.; Goddard, M.E. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS Genet.* **2014**, *10*, e1004198. [CrossRef]
42. Kichaev, G.; Roytman, M.; Johnson, R.; Eskin, E.; Lindström, S.; Kraft, P.; Pasaniuc, B. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **2016**, *33*, 248–255. [CrossRef] [PubMed]
43. Kemper, K.E.; Littlejohn, M.D.; Lopdell, T.; Hayes, B.J.; Bennett, L.E.; Williams, R.P.; Xu, X.Q.; Visscher, P.M.; Carrick, M.J.; Goddard, M.E. Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. *BMC Genom.* **2016**, *17*, 858. [CrossRef] [PubMed]
44. Grisart, B.; Coppieters, W.; Farnir, F.; Karim, L.; Ford, C.; Berzi, P.; Cambisano, N.; Mni, M.; Reid, S.; Simon, P.; et al. 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 Res.* **2002**, *12*, 222–231. [CrossRef]
45. Blott, S.; Kim, J.-J.; Moisio, S.; Schmidt-Küntzel, A.; Cornet, A.; Berzi, P.; Cambisano, N.; Ford, C.; Grisart, B.; Johnson, D.; et al. Molecular dissection of a quantitative trait locus: A phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics* **2003**, *163*, 253–266. [CrossRef]
46. Romanoski, C.E.; Glass, C.K.; Stunnenberg, H.G.; Wilson, L.; Almouzni, G. Roadmap for regulation. *Nature* **2015**, *518*, 314–316. [CrossRef]
47. Moore, J.E.; Purcaro, M.J.; Pratt, H.E.; Epstein, C.B.; Shoresh, N.; Adrian, J.; Kawli, T.; Davis, C.A.; Dobin, A.; Kaul, R.; et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **2020**, *583*, 699–710. [CrossRef]

48. Andersson, L.; Archibald, A.L.; Bottema, C.D.; Brauning, R.; Burgess, S.C.; Burt, D.W.; Casas, E.; Cheng, H.H.; Clarke, L.; Couldrey, C. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* **2015**, *16*, 1–6. [CrossRef]
49. Wang, M.; Hancock, T.P.; Chamberlain, A.J.; Vander Jagt, C.J.; Pryce, J.E.; Cocks, B.G.; Goddard, M.E.; Hayes, B.J. Putative bovine topological association domains and CTCF binding motifs can reduce the search space for causative regulatory variants of complex traits. *BMC Genom.* **2018**, *19*, 395. [CrossRef]
50. Wang, M.; Hancock, T.P.; MacLeod, I.M.; Pryce, J.E.; Cocks, B.G.; Hayes, B.J. Putative enhancer sites in the bovine genome are enriched with variants affecting complex traits. *Genet. Sel. Evol.* **2017**, *49*, 56. [CrossRef]
51. Prowse-Wilkins, C.P.; Wang, J.; Xiang, R.; Garner, J.B.; Goddard, M.E.; Chamberlain, A.J. Putative Causal Variants Are Enriched in Annotated Functional Regions From Six Bovine Tissues. *Front. Genet.* **2021**, *12*, 1027. [CrossRef]
52. Beer, M.A. Predicting enhancer activity and variant impact using gkm-SVM. *Hum. Mutat.* **2017**, *38*, 1251–1258. [CrossRef] [PubMed]
53. van Arensbergen, J.; Pagie, L.; FitzPatrick, V.D.; de Haas, M.; Baltissen, M.P.; Comoglio, F.; van der Weide, R.H.; Teunissen, H.; Võsa, U.; Franke, L.; et al. High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* **2019**, *51*, 1160–1169. [CrossRef] [PubMed]
54. Bulik-Sullivan, B.K.; Loh, P.-R.; Finucane, H.K.; Ripke, S.; Yang, J.; Patterson, N.; Daly, M.J.; Price, A.L.; Neale, B.M.; Schizophrenia Working Group of the Psychiatric Genomics Consortium. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **2015**, *47*, 291–295. [CrossRef] [PubMed]
55. Gusev, A.; Lee, S.H.; Trynka, G.; Finucane, H.; Vilhjálmsson, B.J.; Xu, H.; Zang, C.; Ripke, S.; Bulik-Sullivan, B.; Stahl, E.; et al. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *Am. J. Hum. Genet.* **2014**, *95*, 535–552. [CrossRef]
56. Beer, M.A.; Shigaki, D.; Huangfu, D. Enhancer Predictions and Genome-Wide Regulatory Circuits. *Annu. Rev. Genom. Hum. Genet.* **2020**, *21*, 37–54. [CrossRef]