ABSTRACT

Title of Dissertation:	GENETIC ARCHITECTURE OF COMPLEX TRAITS AND ACCURACY OF GENOMIC SELECTION IN DAIRY CATTLE
	Jicai Jiang, Doctor of Philosophy, 2018
Dissertation directed by:	Assistant Professor Li Ma, Department of Animal and Avian Sciences

Genomic selection has emerged as an effective approach in dairy cattle breeding, in which the key is prediction of genetic merit using dense SNP genotypes, i.e., genomic prediction. To improve the accuracy of genomic prediction, we need better understanding of the genetic architecture of complex traits and more sophisticated statistical modeling. In this dissertation, I developed several computing tools and performed a series of studies to investigate the genetic architecture of complex traits in dairy cattle and to improve genomic prediction models. First, we dissected additive, dominance, and imprinting effects for production, reproduction and health traits in dairy cattle. We found that non-additive effects contributed a non-negligible amount (more for reproduction traits) to the total genetic variance of complex traits in cattle. We also identified a dominant quantitative trait locus (QTL) for milk yield, revealing that detection of QTLs with non-additive effect is possible in genome-wide association studies (GWAS) using a large dataset. Second, we developed a powerful Bayesian method and a fast software tool (BFMAP) for SNP-set association and fine-mapping. We demonstrated that BFMAP achieves a power similar to or higher than existing software tools but is at least a few times faster for association tests. We also showed that BFMAP performs well for fine-mapping and can efficiently integrate fine-mapping with functional enrichment analysis. Third, we performed large-scale GWAS and fine-mapped 35 production, reproduction, and body conformation traits to single-gene resolution. We identified many novel association signals and many promising candidate genes. We also characterized causal effect enrichment patterns for a few functional annotations in dairy cattle genome and showed that our fine-mapping result can be readily used for future functional studies. Fourth, we developed an efficient Bayesian method and a fast computing tool (SSGP) for using functional annotations in genomic prediction. We demonstrated that the method and software have great potential to increase accuracy in genomic prediction and the capability to handle very large data. Collectively, these studies advance our understanding of the genetic architecture of complex traits in dairy cattle and provide fast computing tools for analyzing complex traits and improving genomic prediction.

GENETIC ARCHITECTURE OF COMPLEX TRAITS AND ACCURACY OF GENOMIC SELECTION IN DAIRY CATTLE

by

Jicai Jiang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Assistant Professor Li Ma, Chair
Professor Jiuzhou Song
Dr. Paul VanRaden
Associate Professor Jeffrey O'Connell
Professor Sridhar Hannenhalli

Dedication

To my wife Jie and son Luke, for their love and encouragement

Acknowledgements

I would like to thank my advisor, Dr. Li Ma, for guiding and supporting me over the years. You set an example of excellence as a researcher, mentor, and instructor.

I would like to thank my PhD dissertation committee members for all of their guidance and feedback. It has been particularly a pleasure working with Dr. Paul VanRaden and Dr. Jeffrey O’Connell during my doctoral research.

Special thanks go to my family and friends.

Table of Contents

List of Abbreviations

BF	Bayes Factor
Chr	Chromosome
GEBV	Genomic Estimated Breeding Value
GP	Genomic Prediction
GS	Genomic Selection
GRM	Genomic Relationship Matrix
GWAS	Genome-Wide Association Study
kb	kilo base pairs = 1000 base pairs
LD	Linkage Disequilibrium.
Mb	mega bases pairs = 1000 kb = 1 million base pairs
MCMC	Markov chain Monte Carlo
PIP	Posterior Inclusion Probability
PPC	Posterior Probability of Causality
PTA	Predicted Transmitting Ability
QTL	Quantitative Trait Locus
RMSE	Root-Mean-Square Error
SNP	Single Nucleotide Polymorphism
VB	Variational Bayes
YD	Yield Deviation

Chapter 1: Literature Review

In this chapter, I review recent advances in computational and statistical genetics, biological mechanisms linking a SNP to a trait, and recent advances of statistical methods for genomic prediction. At the end, I propose the questions to address in this dissertation. There are many interesting topics, but only those most relevant to this study are included in this review.

Recent Advances in Computational and Statistical Genetics

Genomic relationship matrix and GREML

Genomic relationship matrix (GRM, often denoted as **G**) is a realized relationship matrix built by using genomic information (basically, whole-genome SNP genotypes). As a relationship matrix, it can be used in most of the scenarios where numerator relationship matrix (often denoted as **A**) is applied. Use of GRM is often straightforward, as there are many well-developed computing techniques involving **A** for genetics and breeding, e.g., variance component estimation and breeding value prediction (Henderson, 1984).

There are several considerations when building a GRM. The first one is the way of using minor allele frequency (MAF). In an early study on GRM, VanRaden (2008) proposed two forms:

$$\mathbf{G} = \mathbf{ZWZ}'/m \qquad (1.1)$$

and

$$\mathbf{G} = \mathbf{ZZ}' \Big/ \sum_j 2p_j(1-p_j)\,, \qquad (1.2)$$

where $\mathbf{Z}$ is a matrix of centered genotypes for additive effects, p_j is the MAF of the jth marker, $\mathbf{W}$ is a diagonal matrix whose jth element is $2p_j(1-p_j)$, and m is the total number of markers. These two formulas, though slightly differing, are actually based on two different assumptions. By using the former one, we assume that all SNPs contribute equally to heritability. In contrast, the assumption for the latter one is that SNPs with a high MAF contribute more to heritability than those with a low MAF. The difference between the assumptions has an impact on estimation of SNP heritability in complex human traits (Speed *et al*, 2017). The second consideration is the way of using linkage disequilibrium (LD) pattern. A few markers can capture causal effects in a high-LD region, while many more markers are needed to do so in a low-LD region (Speed *et al*, 2012). Thus, we may downweight the contribution of SNPs in high-LD regions.

As discussed above, the key to optimizing GRM is weighting SNPs based on MAF and LD pattern. A more recent study generalized the ideas of using MAF and LD pattern and derived an GRM estimator with minimized estimation errors (Wang *et al*, 2017). Basically, the authors obtained the GRM estimation error function whose parameters are weights of markers, and used quadratic programming to obtain the optimal weighting. The method is promising; however, as far as I know, there have not been studies investigating the impacts of using such a GRM on heritability estimation or genomic prediction.

GREML is the use of GRM to estimate variance components via restricted maximum likelihood (REML) (Lee *et al*, 2011; Lee *et al*, 2012; Yang *et al*, 2010). It is a straightforward extension of use of **A** in similar scenarios (Hofer, 1998; Johnson and Thompson, 1995). GREML has been heavily used to investigate the genetic architecture of complex traits. For example, GREML has been used for estimation of SNP heritability (defined as the proportion of phenotypic variation explained by whole-genome SNPs) for many hundreds of traits (Yang *et al,* 2017). GRM has a big effect on heritability estimation by GREML, and different GRM computations (with respect to MAF and LD) may result in a considerable difference in heritability estimates even if the same data set is used (Speed *et al*, 2017). Software tools to address this issue include LDAK (Speed *et al*, 2012) and GCTA (Yang *et al*, 2015; Yang *et al*, 2011a). In addition, GRM is built based on assumptions, so through GREML, we can find which assumption results in better model fitting or prediction. By testing various assumptions, we gain better understanding on the genetic architecture.

Another important use of GREML is heritability partitioning. Multiple GRMs are simultaneously fitted by GREML (namely multi-component GREML), and each GRM is built by a subset of whole-genome SNPs. Accordingly, we can investigate the contribution of each chromosome or each genomic segment to heritability (Yang *et al*, 2011b). SNP grouping by genomic segments also provides an approach for genome-wide association studies (GWAS), namely regional heritability mapping (Caballero *et al*, 2015; Shirali *et al*, 2016). In addition, the SNP grouping can be based on many other genomic features, e.g., MAF bins, LD bins, SnpEff-inferred variant impact (Cingolani *et al*, 2012), so that we let GREML automatically determine the relative importance of each category (Yang *et al*, 2015). Though multi-component GREML is useful for investigating heritability enrichment patterns, it has a limit regarding the number of fitted GRMs. A more general heritability portioning approach has been developed, named stratified LD score regression (Finucane *et al*, 2015), which is also discussed below in this chapter.

GREML can also be used for studying non-additive effects, e.g., dominance, imprinting, and epistasis (Varona *et al*, 2018). Basically, we develop GRMs for these non-additive effects and fit them in multi-component GREML, by which we estimate their separate contributions to phenotypic variation.

Mixed model association methods

GWAS is one of the most commonly used approaches for discovering genetic factors underlying complex traits. It finds SNP-trait associations by tests for whole-genome markers. There have been tens of thousands of unique SNP-trait associations discovered in humans (MacArthur *et al*, 2017) and thousands in livestock species (Hu *et al*, 2016). Over past decade, mixed model association methods have become routine for GWAS, in that they well control population or relatedness structure and are statistically powerful (Yang *et al*, 2014).

Generally, one GRM is needed in GWAS to correct for population structure, so the mixed model used for GWAS can be considered to be a special case of those used in GREML. However, the routine algorithms used for GREML, like AI-REML (Johnson and Thompson, 1995), are infeasible for GWAS, because they may even take hours for only one marker. To make mixed models useful for GWAS, several computing methods have been developed to improve the speed. GRAMMAR is one of the earliest attempts (Aulchenko *et al*, 2007). This approach first fits a null mixed model and then uses the resulting residuals as response variable to perform linear regression on each marker, namely a two-step method. Denote *n* and *m* to be the sample size and the number of markers, respectively. This method has a time complexity of $O(nm)$ (considering only the second step) and is very fast. However, it is usually too conservative. The authors further proposed a method to remedy GRAMMAR, named GRAMMAR-GC (Amin *et al*, 2007). It first computes a genomic deflation factor following the genomic control (GC) method (Devlin and Roeder, 1999), and then divides the GRAMMAR chi-square statistics by the factor. The resulting values are used as new statistics. However, the GC approach still has the same problem.

Routine GREML algorithms have a time complexity of $O(rn^3)$ when one GRM is modeled, where r is the number of iterations required. Kang et al. (2008) reported an eigendecomposition method (named EMMA) to tackle the same problem with a time complexity of $O(n^3 + rn)$. It should be noted that this strategy used for REML had been comprehensively studied by VanRaden in his doctoral dissertation (VanRaden, 1986). The difficulty of evaluating log-likelihood or restricted log-likelihood function is computing inverse and determinant of the variance term (denoted as $\boldsymbol{V}$) and a term involving the inverse and covariate design matrix. Let $\boldsymbol{H} = \boldsymbol{V}/\sigma_e^2 = \lambda\boldsymbol{G} + \boldsymbol{I}$ in which σ_e^2 is the error variance, $\boldsymbol{G}$ is the GRM, and $\lambda = \sigma_g^2/\sigma_e^2$ (σ_g^2 is the genetic variance corresponding to $\boldsymbol{G}$). Let $\boldsymbol{P} = \boldsymbol{SHS}$, where $\boldsymbol{S} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{X}$ is the covariate design matrix. By eigendecompositions of $\boldsymbol{G}$ and $\boldsymbol{P}$, it turns out that log-likelihood or restricted log-likelihood can be formulated as a function of λ. Therefore, the problem is reduced to single-variable optimization. A Wald or likelihood ratio test can accordingly be performed after evaluating the function with respect to λ. Because each marker results in a distinctive matrix $\boldsymbol{X}$ (for covariates and the marker), so EMMA requires an eigendecomposition for each marker in GWAS, leading to a time complexity of $O(mn^3)$. Therefore, EMMA is often impracticable for a medium-size GWAS. Several methods, also based on eigendecomposition, have been developed to address the problem, including GEMMA (Zhou and Stephens, 2012), FaST-LMM (Lippert *et al*, 2011), and MMAP (O'Connell, 2013). Basically, these methods require only one eigendecomposition of GRM (or singular value decomposition of genotype matrix) and have a time complexity of $O(mn^2)$ for association tests.

The aforementioned eigendecomposition methods produce exact test statistics, in the sense that the estimate of λ is unique for each marker. Assuming that the proportion of variance explained by a single SNP is small, the estimate of λ for a model including a SNP will be very similar to that for the null model. This leads us to an approximation method to gain speedup for GWAS. We estimate λ only for the null model, and use it for all association tests. Given λ, we can perform a generalized least squares F-test, resulting in a time complexity of $O(mn^2)$ for association tests. This approximation method has been used in several software tools, e.g., EMMAX (Kang *et al*, 2010), GCTA (Yang *et al*, 2014).

Assume that $\boldsymbol{X}$ is single tested SNP, we can formulate a score test as

$$\chi^2_{\text{score}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\right)^2 \Big/ \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}, \tag{1.3}$$

where **y** is phenotype, and **V** is the same as previously defined. The score test has a time complexity of $O(mn^2)$ for GWAS. GRAMMAR-Gamma further reduce the time complexity to $O(mn)$ by the following approximation (Svishcheva *et al*, 2012):

$$\begin{gathered} \chi^2_{\text{new}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\right)^2 \Big/ \mathbf{X}'\mathbf{X} \\ \chi^2_{\text{score}} = \chi^2_{\text{new}} \big/ \gamma \end{gathered}, \tag{1.4}$$

where $\gamma = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}/\mathbf{X}'\mathbf{X}$. The study suggests that γ is nearly constant across whole-genome markers, and also provides an analytical expression so that γ can be computed before association tests. BOLT-LMM uses similar approximation, but enables a leave-one-chromosome-out (LOCO) approach (Loh *et al*, 2015b). It has a time complexity of $O(mn^{1.5})$. A Bayesian mixture model is also implemented in BOLT-LMM. LOCO SNPs are first fitted in the mixture model, and the resulting residuals are further used as phenotypes in linear regression. This approach is similar to GRAMMAR, but uses LOCO and mixture model to improve statistical power.

In addition, avoiding proximal contamination may be considered (Listgarten *et al*, 2012). Theoretically, excluding markers correlated with the tested SNP in GRM will increase statistical power. The LOCO approach is a special case, which excludes markers on the same chromosome as the tested SNP in GRM. This approach may be especially useful for approximation methods like EMMAX and GCTA.

Though many methods do GWAS in $O(mn^2)$ time, software implementation may make a big difference in real running time. For example, we found MMAP is much faster than EMMAX and GEMMA. In reality, computational speed/convenience is often the most important consideration when deciding which software to use (Eu-Ahsunthornwattana *et al*, 2014).

SNP-set association

SNP-set association test is basically a test for the association between a SNP set and a trait. Therefore, single-marker association test can be considered to be a special case of SNP-set tests. The aforementioned regional heritability mapping by GREML is a type of SNP-set association tests, in the sense that SNPs within each genomic segment form a SNP set. By GREML, we can perform a likelihood ratio test for SNP-set association, which is generally ideal with respect to power. However, this approach is computationally demanding (Cebamanos *et al*, 2014). It has a time complexity of $O(rmn^3)$, where *r* is the number of required iterations, *m* is the number of SNP sets, and *n* is the sample size.

In contrast, score test results in much faster computation, in that it only fits the null model by GREML (or EMMA). Several previous studies have implemented the approach for common variants (Kwee *et al*, 2008; Wu *et al*, 2010) and for rare-variant association testing (namely, SKAT) (Wu *et al*, 2011). SKAT has become one of the most popular software tools for SNP-set association analysis. It can also combine the SKAT score statistic and the weighted burden test statistic (Madsen and Browning, 2009) for rare-variant tests. It should be noted that when the null model is a linear mixed model, it uses EMMA to fit the null. SKAT is fast and has a time complexity of $O(mn^2)$.

It is critical to properly weight SNPs for a SNP-set test. Upweighting a causal variant can improve the power. SKAT authors suggest a weight based on beta probability distribution function, $\sqrt{w_j} = \mathrm{Beta}\left(x = MAF_j; \alpha, \beta\right) \propto x^{\alpha-1}\left(1-x\right)^{\beta-1}$. As shown by the function, Beta(1, 1) gives equal weights to all variants, while Beta(1, 25) upweight rare variants and downweight common variants. When a SNP set contains both rare and common variants, we can set weights for rare and common variants separately (Ionita-Laza *et al*, 2013).

Bayesian fine-mapping

As large-scale sequence data are becoming available, it is now feasible to fine-map a trait to single-variant resolution. Fine-mapping is basically a model selection problem. Many statistical methods have been used to solve the problem, e.g., stepwise selection (Huang *et al*, 2017), exhaustive search limiting maximum model size (Chen *et al*, 2015; Hormozdiari *et al*, 2014; Kichaev *et al*, 2014; Servin and Stephens, 2007), shotgun stochastic search (Benner *et al*, 2016).

Generally, stepwise selection first uses stepwise regression to find independent signals, and then generates a credible variant set for each signal. It is fast and works well for identifying independent causal variants. However, it may fail in some scenarios, e.g., when genotypes of causal variants are highly correlated. Exhaustive search is capable of handling all LD structures; however, it is often infeasible when we aim to find multiple causal effects in many variants (e.g. 1000). We have to limit the maximum model size (usually 3) to reduce the model search burden. Shotgun stochastic search (SSS) overcomes this problem by identifying models with high posterior probability and ignoring models with negligible probability (Hans *et al*, 2007). However, SSS may fail to find all important models for some LD structures, even with a long chain. Additionally, most of the existing fine-mapping tools use summary statistics. Though this is a great feature, direct use of genotypes and phenotypes results in exact computation and is more straightforward, especially in some species where summary statistics is not commonly used (e.g. dairy cattle).

Use of functional annotation is an important topic for fine-mapping. Existing methods (CAVIARBF and PAINTOR) usually use a logistic model, in which a binary variable indicating a variant is causal or not is modeled as response, and categorical functional annotations are used as covariates (Chen *et al*, 2016; Kichaev *et al*, 2014). Such a logistic model is incorporated into a model search scheme. Then, the log-likelihood function is optimized with respect to unknown parameters. This approach often limits the maximum number of causal variants (like 3) and is often impractical for loci containing thousands of variants. In addition, the model search results for a function annotation cannot be re-used for other functional annotations, further increasing the computational burden.

LD score regression

LD score regression (LDSC) was proposed very recently, which addresses the use of GWAS summary statistics for estimating SNP heritability (Bulik-Sullivan *et al*, 2015b), partitioning heritability (Finucane *et al*, 2015), and estimating genetic correlation between traits (Bulik-Sullivan *et al*, 2015a). Define the LD score of variant *j* as

$$l_j := \sum\nolimits_k r_{jk}^2 , \tag{1.5}$$

in which r_{jk} is the genotype correlation between variants *j* and *k*. When there is no confounding (or inflation) in summary statistics, we can get

$$\mathrm{E}\left[\chi_j^2\right] = \frac{Nh_g^2}{M} l_j + 1, \tag{1.6}$$

in which N is the sample size, M is the number of markers, and h_g^2 is the SNP heritability. This formula leads us to a linear regression of summary statistics on LD score, which produces a heritability estimate. Furthermore, LD score of variant *j* can be partitioned into multiple parts based on functional annotations. For category *c*, the LD score is computed as follows

$$l_j\left(j,c\right) := \sum_k a_{ck} r_{jk}^2\,, \tag{1.7}$$

in which a_{ck} (0 or 1) is a variable indicating whether variant *k* belongs to category *c*. Similar to equation (1.6), it turns out that

$$\mathrm{E}\left[\chi_j^2\right] = N \sum_c l(j,c)\tau_c + 1, \tag{1.8}$$

where τ_c is a heritability-related term for category *c*. LD score regression is thus conceived, which is named stratified LD score regression (S-LDSC) (Finucane *et al*, 2015). After obtaining the estimate of τ_c in S-LDSC, the heritability explained by variants in category *c* is readily computed. In addition, LDSC can also be used to estimate genetic correlation between traits (Bulik-Sullivan *et al*, 2015a). In the multi-trait LD score regression, summary statistics from linear regression are sufficient; that is, mixed model is not necessarily needed. This feature makes it especially useful for large-scale GWAS data.

Note that LDSC is based on the infinitesimal model for complex traits. S-LDSC estimates enrichment of heritability for functional annotations. In contrast, incorporation of function annotation in fine-mapping is based on a sparse model, and the resulting estimate is actually an enrichment of causal variants. Despite the difference in definition, the two types of enrichments may have similar estimates (Sveinbjornsson *et al*, 2016).

Biological Mechanisms Linking a SNP to a Trait

Various mechanisms

Protein-coding genes contain many elements, e.g., enhancer/silencer, promoter, 5' untranslated region (UTR), extron, intron and 3' UTR (Lewin, 2008). A SNP can possibly be in any of these elements. Depending on position and function, SNPs can be grouped into more than 10 categories (e.g., stop gained, stop lost, splice acceptor, splice donor, missense, synonymous, etc.), which has been clearly defined by Sequence Ontology (Cunningham *et al*, 2015). As reported in previous studies, many of these types of SNPs can be causal variants for a trait. Here, we provide an incomplete review on various biological mechanisms linking a SNP to a trait. Actually, the effect of a SNP on a trait has to rely on its effect on corresponding protein. Thus, this review is centered on classifying how a SNP can change corresponding protein product in terms of structure, stability, activity or expression.

Missense mutations cause protein sequence changes which further result in changes of protein stability and/or enzyme activity. One typical example is the *DGAT1 K232A* quantitative trait nucleotide which affects milk yield and composition in dairy cattle (Grisart *et al*, 2004). The *DGAT1* gene encodes an enzyme, diglyceride acyltransferase (DGAT). The enzyme encoded by the *K* allele has significantly higher activity than encoded by the *A* allele. As a result, the *A* to *K* substitution effect has been shown to correspond to ~0.35% of milk fat percentage, and to ~10 kg of milk fat in the Holstein dairy cattle (Grisart *et al*, 2004).

Nonsense mutations can have higher impact on protein products than missense mutations, because it results in a premature stop codon in gene sequence and in truncated, incomplete protein product which usually loses its function. Nonsense mutations may cause many diseases (Mendell and Dietz, 2001). For example, either *G542X* or *W1282X* mutation in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene can cause cystic fibrosis, because either of the nonsense mutations results in nonfunctional CFTR protein product (O'Sullivan and Freedman, 2009).

Splice site mutations can result in non-functional or abnormal protein products, because they may induce remaining of introns or missing of exons in mature mRNA during transcript processing. Alavi *et al* (2007) reported a mouse model carrying a splice site mutation in the *Opa1* gene encoding GTPase. The mutation induces a skipping of exon 10 and leads to an in-frame deletion of 27 amino acid residues in the GTPase domain. Their study showed homozygous mutant mice die in utero, and heterozygous mutants are viable and of normal habitus but exhibit clear symptoms of optic atrophy.

SNPs in gene regulatory regions can also be causal mutation for a trait, as they may affect regulatory elements (e.g., promoter, enhancer/silencer) and thus gene expression. For example, De Gobbi *et al* (2006) found that a regulatory SNP (rSNP) can cause a human genetic disease, inherited blood disorder alpha thalassemia, by creating a new transcriptional promoter. The rSNP is in a noncoding region between the alpha-globin genes and their upstream regulatory elements, and it can create a new promoter-like element that interferes with normal activation of all downstream alpha-like globin genes, which significantly down-regulated gene expression.

A SNP in regulatory element can even affect transcription of a gene that is distant from it. The underlying mechanism is that distant regulatory elements can physically interact with promoters of target genes by long range chromatin loops (Dean, 2011). For example, Zhang *et al* (2012) found an enhancer formed a 1Mb chromatin loop to the *SOX9* gene. They reported that two SNPs in the enhancer can affect enhancer activity and thus impose allele-specific expression of *SOX9* which is further associated with prostate cancer.

SNPs in microRNA (miRNA) binding sites in the 3' UTRs of target genes can affect phenotype of a trait through modulating the regulatory loop between miRNAs and their target genes (Zhang *et al*, 2011). MiRNAs are single-stranded, noncoding RNA molecules, involving in many biological processes (Ambros, 2004). Most miRNAs bind to target sequences located within the 3' UTR of mRNAs by base pairing, resulting in the cleavage of target mRNAs or repression of their translation (Meister and Tuschl, 2004). For instance, ryanodine receptor 3 gene (*RYR3*), which is important for the growth, morphology and migration of breast cancer cells, contains a putative binding site for microRNA-367 (miR-367) in its 3' UTR. There is an A/G SNP (rs1044129) located in the miR-367 binding site. Zhang *et al* (2011) reported miR-367 has a higher binding affinity for the A genotype than for the G genotype. Higher binding affinity results in lower gene expression. The expression of *RYR3* is thus affected by the SNP genotype and rs1044129 is a unique SNP that resides in a miRNA-gene regulatory loop that affects breast cancer risk (Zhang *et al*, 2011).

Identifying associations

Here, we focus on discussing complex traits or diseases rather than simple Mendelian traits. Human diseases and agricultural traits of interest are usually complex traits with a highly polygenic architecture. QTLs usually have very small effect, thus we need very large sample size to detect the effect. It is more difficult to detect the effects of QTLs with low MAF or rare QTLs, even though rare QTLs may probably have relatively large effects. In addition, single-SNP analysis must rely on substantial LD between markers and QTL. Some QTLs may not be well tagged by markers, making it harder to find the SNP-trait associations. Due to these challenges, single-SNP genome-wide association study (GWAS) can usually find a very limited number of SNP-trait associations even with a sample of >100,000 individuals (Yang *et al*, 2012).

Set tests can usually increase the power of finding SNP-trait associations by using a set of SNPs together for a test. For example, a recent study using regional heritability mapping method reported that ≥71% of 1-Mb genomic regions harbor ≥1 variant influencing schizophrenia risk, that is, >2,000 causal variants (Loh *et al*, 2015a), which is much powerful than single-SNP tests. However, this approach often compromises on resolution.

Ascertaining mechanistic links

Ascertaining the mechanistic links for SNP-trait associations is often difficult, because this implies we need to find causal variants. A common, feasible approach is post-GWAS prioritization; that is, we use available information (e.g., gene annotation, variant effect prediction, documented studies, eQTL data, etc.) to prioritize the GWAS results (Hou and Zhao, 2013). For a locus of interest, however, we may find a number of candidate variants that have distinct mechanistic links. We have to analyze these SNPs one by one to determine which is causal and find the mechanistic link. For instance, when we find a region of interest which covers several genes, we may find a number of missense SNPs, miRNA binding site SNPs and promoter SNPs, separately. To determine which SNPs are causal, we need to use other information, e.g., variant effect prediction for missense SNPs and gene expressions, etc. Considering tissue-specific gene expression, we have to investigate gene expressions in various tissues. Otherwise, we can only obtain a general profile for mechanisms (Nicolae *et al*, 2010; Pal *et al*, 2015) rather than accurately ascertain the mechanistic link for a causal SNP.

A more promising direction is integrating GWAS results with other types of data (e.g., gene expression) to ascertain specific mechanistic links. (Zhu *et al*, 2016) propose a method (SMR) that integrates summary-level data from GWAS with data from eQTL studies to identify genes whose expression levels are associated with a complex trait. Their method can find links between SNPs, gene expression and trait, but only work on cis-regulatory SNPs. Such integrations have a few challenges. Compared to GWAS summary data or SNP genotype data, gene expression data are relatively lacking. In addition, both eQTL analysis and GWAS impose multiple comparisons problem, such that the information we gain from their integration may be limited.

Recent Advances of Statistical Methods for Genomic Prediction

Nowadays, genomic prediction (GP) is widely used in plant and animal breeding programs and has been well proven to be effective (Garcia-Ruiz *et al*, 2016). Since the seminal work of Meuwissen *et al* (2001) for predicting genomic breeding values in animal and plant breeding, a number of genomic prediction methods have been developed and extensively investigated based on different algorithms, e.g., semi-parametric methods (Gianola *et al*, 2006), nonlinear regression (VanRaden, 2008), Bayesian LASSO (Legarra *et al*, 2011), and the Bayesian alphabet (Gianola, 2013; Habier *et al*, 2011). Generally, these parametric methods assume that the effect of each marker is independently distributed with a specific prior distribution given by corresponding statistical methods. Clearly, such an assumption of independent distribution for each SNP effect is statistically inappropriate, especially when the adjacent markers are in high LD with the same causal gene. This unrealistic assumption potentially sacrifices the prediction accuracy to some extent. To address this issue, Yang and Tempelman (2011) proposed a first-order antedependence model to account for the nonstationary correlations between SNP markers through assuming a linear relationship between the effects of adjacent markers. As expected, the proposed antedependence-based GP models outperformed their conventional counterparts in the prediction accuracy of genomic merit in the context of single-trait analyses.

Many of the aforementioned methods have been extended to joint prediction of multiple traits, e.g., BayesA, BayesC (Calus and Veerkamp, 2011), BayesCπ (Jia and Jannink, 2012), antedependence-based BayesA (Jiang *et al*, 2015). These methods assume that that a locus simultaneously affects all the traits or none of them in the analysis. To relax this assumption, Cheng et al. recently proposed more general multi-trait BayesCπ and BayesB methods allowing a broader range of mixture priors (Cheng *et al*, 2018).

All the methods mentioned above generally assume that the reference population is genotyped and phenotyped. To make use of both genotypes and pedigree information, single-step genomic BLUP was proposed, in which a relationship matrix combining genomic relationships and pedigree relationships is constructed (Legarra *et al*, 2014). Much work has been done to improve the stability, efficiency and flexibility of the method (Lourenco *et al*, 2017; Masuda *et al*, 2016; Misztal *et al*, 2013a; Misztal *et al*, 2013b). Another promising way of enhancing genomic prediction is to incorporate existing biological information into GP models. A previous study shows that using gene annotation can produce higher prediction accuracy for some traits (Gao *et al*, 2017).

Computational efficiency is critical for GP methods, because the reference population in plant and animal breeding programs is quickly growing. For example, there have been more than 2.6 million genotyped animals in genomic evaluation at the Council on Dairy Cattle Breeding (CDCB) (https://queries.uscdcb.com/Genotype/counts.html). Many Bayesian methods are based on Markov chain Monte Carlo (MCMC), a time-demanding algorithm. Though they may be theoretically advantageous and perform well for small data, it is impractical to use them in real-world applications. This is why GBLUP and non-linear regression (VanRaden, 2008) are favored in practice. Nevertheless, there are several

MCMC-based computing tools which are well optimized and capable of processing large data sets, e.g., BayesRv2 (Moser *et al*, 2015).

Specific Aims

The overall objective of this study is to gain knowledge on the genetic architecture of complex traits and to develop a method for using the knowledge to improve genomic prediction in dairy cattle.

Aim 1: Dissect additive and non-additive genetic effects for production, reproduction and health traits in dairy cattle.

Aim 2: Develop a powerful method and a fast software tool for SNP-set association and fine-mapping.

Aim 3: Identify QTLs underlying the complex traits in Holstein cattle using imputed sequence data, and fine-map the traits to single-gene resolution.

Aim 4: Develop an efficient method and a fast computing tool for using functional annotations in GS.

References

Alavi MV, Bette S, Schimpf S, Schuettauf F, Schraermeyer U, Wehrl HF *et al* (2007). A splice site mutation in the murine Opa1 gene features pathology of autosomal dominant optic atrophy. *Brain* **130**(Pt 4)**:** 1029-1042.

Ambros V (2004). The functions of animal microRNAs. *Nature* **431**(7006)**:** 350-355.

Amin N, van Duijn CM, Aulchenko YS (2007). A genomic background based method for association analysis in related individuals. *PLoS One* **2**(12)**:** e1274.

Aulchenko YS, de Koning DJ, Haley C (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**(1)**:** 577-585.

Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**(10)**:** 1493-1501.

Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR *et al* (2015a). An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**(11)**:** 1236-1241.

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C *et al* (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**(3)**:** 291-295.

Caballero A, Tenesa A, Keightley PD (2015). The nature of genetic variation for complex traits revealed by GWAS and regional heritability mapping analyses. *Genetics***:** genetics. 115.177220.

Calus MP, Veerkamp RF (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution* **43**(1)**:** 26.

Cebamanos L, Gray A, Stewart I, Tenesa A (2014). Regional heritability advanced complex trait analysis for GPU and traditional parallel architectures. *Bioinformatics* **30**(8)**:** 1177-1179.

Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA *et al* (2015). Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**(3)**:** 719-736.

Chen W, McDonnell SK, Thibodeau SN, Tillmans LS, Schaid DJ (2016). Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics. *Genetics* **204**(3)**:** 933-958.

Cheng H, Kizilkaya K, Zeng J, Garrick D, Fernando R (2018). Genomic Prediction from Multiple-Trait Bayesian Regression Methods Using Mixture Priors. *Genetics* **209**(1)**:** 89-103.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L *et al* (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**(2)**:** 80-92.

Cunningham F, Moore B, Ruiz-Schultz N, Ritchie GR, Eilbeck K (2015). Improving the Sequence Ontology terminology for genomic variant annotation. *J Biomed Semantics* **6:** 32.

De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H *et al* (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **312**(5777)**:** 1215-1217.

Dean A (2011). In the loop: long range chromatin interactions and gene regulation. *Brief Funct Genomics* **10**(1)**:** 3-10.

Devlin B, Roeder K (1999). Genomic control for association studies. *Biometrics* **55**(4)**:** 997-1004.

Eu-Ahsunthornwattana J, Howey RA, Cordell HJ (2014). Accounting for relatedness in family-based association studies: application to Genetic Analysis Workshop 18 data. *BMC Proc* **8**(Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo)**:** S79.

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR *et al* (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**(11)**:** 1228-1235.

Gao N, Martini JWR, Zhang Z, Yuan X, Zhang H, Simianer H *et al* (2017). Incorporating Gene Annotation into Genomic Prediction of Complex Phenotypes. *Genetics* **207**(2)**:** 489-501.

Garcia-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-Lopez FJ, Van Tassell CP (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc Natl Acad Sci U S A* **113**(28)**:** E3995-4004.

Gianola D (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics***:** genetics. 113.151753.

Gianola D, Fernando RL, Stella A (2006). Genomic assisted prediction of genetic value with semi-parametric procedures. *Genetics*.

Grisart B, Farnir F, Karim L, Cambisano N, Kim JJ, Kvasz A *et al* (2004). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc Natl Acad Sci U S A* **101**(8)**:** 2398-2403.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011). Extension of the Bayesian alphabet for genomic selection. *BMC bioinformatics* **12**(1)**:** 186.

Hans C, Dobra A, West M (2007). Shotgun Stochastic search for "Large p" regression. *Journal of the American Statistical Association* **102**(478)**:** 507-516.

Henderson CR (1984). *Applications of linear models in animal breeding*. Guelph : University of Guelph.

Hofer A (1998). Variance component estimation in animal breeding: a review. *Journal of Animal Breeding and Genetics* **115**(1-6)**:** 247-265.

Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* **198**(2)**:** 497-508.

Hou L, Zhao H (2013). A review of post-GWAS prioritization approaches. *Front Genet* **4:** 280.

Hu ZL, Park CA, Reecy JM (2016). Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Res* **44**(D1)**:** D827-833.

Huang H, Fang M, Jostins L, Umicevic Mirkov M, Boucher G, Anderson CA *et al* (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**(7662)**:** 173-178.

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X (2013). Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet* **92**(6)**:** 841-853.

Jia Y, Jannink J-L (2012). Multiple trait genomic selection methods increase genetic value prediction accuracy. *Genetics***:** genetics. 112.144246.

Jiang J, Zhang Q, Ma L, Li J, Wang Z, Liu J (2015). Joint prediction of multiple quantitative traits using a Bayesian multivariate antedependence model. *Heredity* **115**(1)**:** 29.

Johnson D, Thompson R (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *Journal of dairy science* **78**(2)**:** 449-456.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB *et al* (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**(4)**:** 348-354.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ *et al* (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**(3)**:** 1709-1723.

Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL *et al* (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**(10)**:** e1004722.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP (2008). A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet* **82**(2)**:** 386-397.

Lee SH, Wray NR, Goddard ME, Visscher PM (2011). Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**(3)**:** 294-305.

Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**(19)**:** 2540-2542.

Legarra A, Christensen OF, Aguilar I, Misztal I (2014). Single Step, a general approach for genomic selection. *Livestock Science* **166:** 54-65.

Legarra A, Robert-Granié C, Croiseau P, Guillaume F, Fritz S (2011). Improved Lasso for genomic selection. *Genetics research* **93**(1)**:** 77-87.

Lewin B (2008). *Genes 9*. Jones & Bartlett Learning.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011). FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**(10)**:** 833-835.

Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D (2012). Improved linear mixed models for genome-wide association studies. *Nat Methods* **9**(6)**:** 525-526.

Loh PR, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ *et al* (2015a). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet* **47**(12)**:** 1385-1392.

Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM *et al* (2015b). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**(3)**:** 284-290.

Lourenco DAL, Fragomeni BO, Bradford HL, Menezes IR, Ferraz JBS, Aguilar I *et al* (2017). Implications of SNP weighting on single-step genomic predictions for different reference population sizes. *J Anim Breed Genet* **134**(6)**:** 463-471.

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E *et al* (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**(D1)**:** D896-D901.

Madsen BE, Browning SR (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**(2)**:** e1000384.

Masuda Y, Misztal I, Tsuruta S, Legarra A, Aguilar I, Lourenco DAL *et al* (2016). Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. *J Dairy Sci* **99**(3)**:** 1968-1974.

Meister G, Tuschl T (2004). Mechanisms of gene silencing by double-stranded RNA. *Nature* **431**(7006)**:** 343-349.

Mendell JT, Dietz HC (2001). When the message goes awry: disease-producing mutations that influence mRNA content and performance. *Cell* **107**(4)**:** 411-414.

Meuwissen TH, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4)**:** 1819-1829.

Misztal I, Aggrey SE, Muir WM (2013a). Experiences with a single-step genome evaluation. *Poult Sci* **92**(9)**:** 2530-2534.

Misztal I, Tsuruta S, Aguilar I, Legarra A, VanRaden PM, Lawlor TJ (2013b). Methods to approximate reliabilities in single-step genomic evaluation. *J Dairy Sci* **96**(1)**:** 647-654.

Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM (2015). Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *Plos Genetics* **11**(4).

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**(4)**:** e1000888.

O'Connell JR. (2013). *63th Annual Meeting of The American Society of Human Genetics*.

O'Sullivan BP, Freedman SD (2009). Cystic fibrosis. *Lancet* **373**(9678)**:** 1891-1904.

Pal LR, Yu CH, Mount SM, Moult J (2015). Insights from GWAS: emerging landscape of mechanisms underlying complex trait disease. *BMC Genomics* **16 Suppl 8:** S4.

Servin B, Stephens M (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**(7)**:** e114.

Shirali M, Pong-Wong R, Navarro P, Knott S, Hayward C, Vitart V *et al* (2016). Regional heritability mapping method helps explain missing heritability of blood lipid traits in isolated populations. *Heredity (Edinb)* **116**(3)**:** 333-338.

Speed D, Cai N, Consortium U, Johnson MR, Nejentsev S, Balding DJ (2017). Reevaluation of SNP heritability in complex human traits. *Nat Genet* **49**(7)**:** 986-992.

Speed D, Hemani G, Johnson MR, Balding DJ (2012). Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* **91**(6)**:** 1011-1021.

Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A, Masson G *et al* (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet* **48**(3)**:** 314-317.

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012). Rapid variance components-based method for whole-genome association analysis. *Nat Genet* **44**(10)**:** 1166-1170.

VanRaden PM (1986). Computational strategies for estimation of variance components. *Retrospective Theses and Dissertations*. 8319.

VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**(11)**:** 4414-4423.

Varona L, Legarra A, Toro MA, Vitezica ZG (2018). Non-additive Effects in Genomic Selection. *Front Genet* **9:** 78.

Wang B, Sverdlov S, Thompson E (2017). Efficient Estimation of Realized Kinship from Single Nucleotide Polymorphism Genotypes. *Genetics* **205**(3)**:** 1063-1078.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ *et al* (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* **86**(6)**:** 929-942.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**(1)**:** 82-93.

Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH *et al* (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47**(10)**:** 1114-1120.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR *et al* (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**(7)**:** 565-569.

Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, Replication DIG *et al* (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**(4)**:** 369-375, S361-363.

Yang J, Lee SH, Goddard ME, Visscher PM (2011a). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**(1)**:** 76-82.

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM *et al* (2011b). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* **43**(6)**:** 519-525.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**(2)**:** 100-106.

Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet* **49**(9)**:** 1304-1310.

Yang W, Tempelman RJ (2011). A Bayesian antedependence model for whole genome prediction. *Genetics***:** genetics. 111.131540.

Zhang L, Liu Y, Song F, Zheng H, Hu L, Lu H *et al* (2011). Functional SNP in the microRNA-367 binding site in the 3'UTR of the calcium channel ryanodine receptor gene 3 (RYR3) affects breast cancer risk and calcification. *Proc Natl Acad Sci U S A* **108**(33)**:** 13653-13658.

Zhang X, Cowper-Sal lari R, Bailey SD, Moore JH, Lupien M (2012). Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. *Genome Res* **22**(8)**:** 1437-1446.

Zhou X, Stephens M (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**(7)**:** 821-824.

Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE *et al* (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**(5)**:** 481-487.

Chapter 2: Dissection of Additive, Dominance, and Imprinting Effects for Production and Reproduction Traits in Holstein Cattle

Abstract

Background: Although genome-wide association and genomic selection studies have primarily focused on additive effects, dominance and imprinting effects play an important role in mammalian biology and development. The degree to which these non-additive genetic effects contribute to phenotypic variation and whether QTL acting in a non-additive manner can be detected in genetic association studies remain controversial.

Results: To empirically answer these questions, we analyzed a large cattle dataset that consisted of 42,701 genotyped Holstein cows with genotyped parents and phenotypic records for eight production and reproduction traits. SNP genotypes were phased in pedigree to determine the parent-of-origin of alleles, and a three-component GREML was applied to obtain variance decomposition for additive, dominance, and imprinting effects. The results showed a significant non-zero contribution from dominance to production traits but not to reproduction traits. Imprinting effects significantly contributed to both production and reproduction traits. Interestingly, imprinting effects contributed more to reproduction traits than to production traits. Using GWAS and imputation-based fine-mapping analyses, we identified and validated a dominance association signal with milk yield near *RUNX2*, a candidate gene that has been associated with milk production in mice. When adding non-additive effects into the prediction models, however, we observed little or no increase in prediction accuracy for the eight traits analyzed.

Conclusions: Collectively, our results suggested that non-additive effects contributed a non-negligible amount (more for reproduction traits) to the total genetic variance of complex traits in cattle, and detection of QTLs with non-additive effect is possible in GWAS using a large dataset.

Keywords: Variance Decomposition, Additive, Dominance, Imprinting, Cattle, Dairy Traits, QTL

Introduction

Both dominance and imprinting play an important role in mammalian biology and development (Moore and Haig, 1991). Though one may naturally assume that dominance and imprinting effects affect economically important traits in plants and animals, it remains controversial how much phenotypic variation can be attributed to these non-additive effects, how many quantitative trait loci (QTL) follow non-additive inheritance, and whether incorporating non-additive genetic effects will benefit genomic prediction (Carlborg and Haley, 2004; Hill *et al*, 2008; Manolio *et al*, 2009). Generally, contribution of non-additive genetic effects varies for different types of traits. For example, genetic variation associated with fitness-related traits is due mostly to low frequency, deleterious variants, so these traits typically show relatively high non-additive variance out of the total genetic variation (Hill *et al*, 2008).

Several studies have been conducted to decompose dominance genetic effects from the total genetic variance of complex traits, theoretically (Da *et al*, 2014; Su *et al*, 2012; Vitezica *et al*, 2013; Wang *et al*, 2014) and empirically (Aliloo *et al*, 2016b; Sun *et al*, 2014; Wittenburg *et al*, 2015; Xiang *et al*, 2016). A few recent studies have tried to add imprinting effects into the decomposition of total genetic variation (Guo *et al*, 2016; Hu *et al*, 2015; Lopes *et al*, 2015; Nishio and Satoh, 2015). These studies indicated that non-additive effects have a significant contribution to the total genetic variance, but it is still questionable whether or not this contribution can be robustly translated into more accurate genomic prediction in real populations. More recently, it was shown that mating programs increased rates of genetic gain when non-additive genetic effects were included (Aliloo *et al*, 2016a; Sun *et al*, 2013; VanRaden, 2016a).

Further understanding of the contribution of non-additive effects to the genomic prediction and mating allocation programs will benefit livestock production in the long term.

Gene mapping studies have primarily focused on genetic variants with additive effects. Although many empirical studies have reported non-negligible contributions from non-additive effects to complex traits, QTLs with non-additive effects are still difficult to identify in animal and human gene mapping studies, largely due to the low statistical power in the testing for non-additive effects of individual loci (Ma *et al*, 2012). The large dairy genomics database maintained by the Council on Dairy Cattle Breeding (CDCB) and the USDA Animal Genomics and Improvement Laboratory (AGIL; Beltsville, MD) represents a powerful dataset for mapping QTLs with non-additive effects.

To empirically address questions related to dominance and imprinting effects of complex traits, we analyzed a large cattle dataset that consisted of more than 40K Holstein cows with SNP genotypes, pedigree information, and eight yield deviation (YD) phenotypes (milk yield, fat yield, protein yield, daughter pregnancy rate, cow conception rate, heifer conception rate, somatic cell score, and productive life). Both parents of these cows were also genotyped to phase the parental inheritance of SNPs of the cows. The aims of this study were to estimate the relative contribution of additive, dominance, and imprinting effects to dairy production and reproduction traits, to identify QTLs with dominance or imprinting effects, and to investigate whether adding these non-additive genetic components improves the prediction accuracy of genomic selection in real data.

Results

Variance decomposition of additive, dominance, and imprinting effects

Using 42,701 Holstein cows with YD phenotypes, SNP genotypes, and two genotyped parents, we decomposed the total genetic value of eight dairy traits into additive, dominance, and imprinting effects, estimating corresponding variance components (Table 2.1). For the eight traits analyzed, the number of animals with YD phenotype ranged from 12,911 (productive life) to 29,811 (milk, fat, and protein yields). Overall, production traits (milk, fat, and protein yields) exhibited a different pattern from reproduction traits (daughter pregnancy, cow conception, and heifer conception rates). As shown in Table 2.1, the broad-sense heritability (H^2 = proportion of total genetic variance in phenotypic variance) was 31.9-38.6% for production traits and 1.4-7.9% for reproduction traits, respectively. The narrow-sense heritability (h^2 = proportion of additive genetic variance in phenotypic variance) was 27.2-33.8% for production traits and only 0.8-5.1% for reproduction traits, respectively. Proportions of dominance variance in phenotypic variance were significantly higher ($P < 0.05$) for production traits (2.5%-4.0%) than for reproduction traits (0.2%-1.1%), but the proportions in total genetic variance are higher for reproduction traits. The variance explained by imprinting effect was very low for all eight traits, <1% of the phenotypic variance for production traits and 1-2% for reproduction traits. However, these imprinting effects were significantly larger than zero for most production and reproduction traits ($P < 0.05$). Moreover, for reproduction traits that have a low heritability, imprinting effects explained a relatively large portion of the total genetic variance (20.9% for daughter pregnancy rate, 26.4% for cow conception rate, and 35.4% for heifer conception rate), which were significantly higher than those for production traits (P<0.05).

For comparison purposes, the total genetic variance was decomposed into the genotypic imprinting value plus either breeding value and dominance deviation using a classical model that considered allele frequencies (Vitezica *et al*, 2013) or additive and dominance effects that did not consider allele frequencies (see Materials and Methods). As shown in Table 2.2, results from these two decomposition models were consistent. It is worth noting that estimated H^2 from the two models was exactly the same for all eight traits. In addition, the proportion of variance explained by imprinting effects was the same for the two models. These results were consistent with theoretical expectations (Álvarez-Castro, 2015; Vitezica *et al*, 2013). In theory, the two variance decomposition models are equivalent to each other with the same predicted phenotypic values and residuals. First, the sum of additive and dominance genetic variances is equal to the sum of the variances of breeding value and dominance deviation, under a few common assumptions (see Materials and Methods). With a stronger condition, the sum of individual breeding value and dominance deviation will be equal to the sum of individual genotypic additive and dominance values. Second, individual genotypic imprinting values of the two models are the same, asserting an equivalence of imprinting variance components. We observed all of these results across all eight traits, as shown in Fig. 2.1 for milk (other traits have the same pattern). Additionally, we confirmed that individual residual estimates of the two models are the same (see the right panels in Fig 2.1).

Genomic relationship matrix (GRM) based variance decomposition is highly dependent on the assumption of polygenic genetic architecture, as genome-wide SNP genotypes are used with equal weights. Existing GWAS have provided evidence of a polygenic architecture of additive effects in most complex traits (Kemper and Goddard, 2012). However, we have no such knowledge for dominance and imprinting effects. To investigate the influence of this polygenic assumption on variance components estimation, we performed simulations to determine if our models have biases when there are only a few dominance or imprinting QTLs. Simulation results showed that GREML could accurately estimate variances for genotypic dominance and imprinting values for a moderate-heritability trait like milk yield, even when only 10 dominance and imprinting QTLs were simulated for a trait with polygenic additive effects, respectively (Fig. 2.2A). For a low-heritability trait like daughter pregnancy rate, GREML also performed well for both lowly and highly polygenic architectures of dominance and imprinting effects (Fig. 2.2B). Using simulation, we demonstrated the robustness of our approach to the assumption of polygenic genetic architecture.

Genome-wide association study of dominance and imprinting effects

We performed a whole-genome single-marker scan for additive, dominance, and imprinting effects on all eight traits. To increase computational efficiency, we used a two-step approach to remove polygenic effects from the data: 1) a mixed model with genomic relationship matrices to generate residuals; followed by, 2) a GWAS scan using residuals from the mixed model as the phenotype. Although our two-step strategy has slightly lower power than a single-step mixed model, we identified a novel dominance signal on chromosome 23 that was associated with milk yield (Fig. 2.3). We then used a single-step mixed model to re-analyze the SNPs near the dominance signal, generating appropriate results for the associated SNPs (Table 2.3). The top 2 SNPs, Hapmap48809-BTA-55698 and BovineHD2300004730, showed a strong dominance association with milk yield with $P = 9.54 \times 10^{-8}$ and $P = 6.33 \times 10^{-8}$, respectively. BovineHD2300004730 is 71 kb upstream of the *RUNX2* gene. The *RUNX2* gene has been previously reported to be a novel regulator of mammary epithelial cell fate in development and breast cancer, and it has also been shown that exogenous transgenic expression of *RUNX2* in mammary epithelial cells blocked milk production (Owens *et al*, 2014).

We further used an independent validation data set consisting of ~5,500 younger cows with both genotypes and milk yield phenotypes, which were collected after the initial analysis, to validate the dominance signal associated with milk yield. A mixed-model based method was used to test the association between milk yield and 50 SNPs around the peak signal. This validation analysis provided clear statistical evidence for the dominance association at BovineHD2300004730 with milk yield ($P = 7.41 \times 10^{-4}$; Fig. 2.4). Additionally, we found that the dominance effect was slightly larger than the additive effect at BovineHD2300004730 in both the discovery and validation data sets, suggesting complete dominance or even over-dominance inheritance of the underlying QTL.

We found no other significant non-additive effects for any trait using a genome-wide significance level of 1×10^{-6}. Nevertheless, there were a few nominally significant peaks for dominance or imprinting effects shown in the Manhattan plots, such as the peak for imprinting effect on chromosome 6 for somatic cell score the one at the end of chromosome 10 for cow conception rate. Since a one-step mixed model is more powerful than a two-step scan, we selected 10 nominally significant non-additive association signals and used a one-step mixed-model to test the associations for the top three SNPs within each peak. This one-step re-analysis found a genome-wide significant dominance association on chromosome 10 with both fat and protein yields. However, this dominance signal was not confirmed in the validation data set.

Fine-mapping of the dominance GWAS peak near *RUNX2*

From our GWAS and validation analyses, we selected BovineHD2300004730 (Chr23:18,600,456) as our target region for fine-mapping using sequence-based imputation. Based on the LD decay pattern between BovineHD2300004730 and nearby variants derived from the sequences of 443 Holstein bulls from the 1000 Bull Genomes project (Run 5.0) (Daetwyler *et al*, 2014), we chose the region of ±500 kb from the targeted SNP for fine mapping to cover all the variants with a LD level of $r^2 > 0.2$ with BovineHD2300004730 (Fig. 2.5A). Using the 443 Holstein sequences as reference, we then imputed sequence-level SNPs in the targeted region for 29,811 cows. After post-imputation quality control, a total of 652 variants were included in a two-step association analysis for milk yield.

The fine-mapping study identified 38 imputed variants with a stronger association than BovineHD2300004730 (Fig. 2.5B). The smallest *P*-value for dominance effect (8.64×10^{-9}) was found at two variants, one in the first intron of *RUNX2* (Chr23:18676057) and the other between *SUPT3H* and *RUNX2* (Fig. 2.5B). Although the 38 variants were all modifiers, the fine mapping analysis provided more evidence that the QTL is close to the *RUNX2* gene. Additionally, most of the variants had a larger dominance effect than additive effect, which was consistent with our original results supporting a dominant or over-dominant mode of inheritance. To investigate whether or not the significant associations were resulted from a single signal, we conducted a conditional analysis by adding the top variant (Chr23:18676057) as a covariate into the association test of each of the remaining 651 variants. This analysis revealed that the significant additive associations disappeared while the dominance signals remained (Fig. 2.6A). Conditioning on both the additive and the dominance effects eliminated all of the significant additive and dominance associations, indicating a single underlying QTL responsible for the association (Fig. 2.6B).

Since we imputed relatively low-density genotypes to sequence genotypes, imputation accuracy was a concern because poor imputation may result in smaller *P*-values in our fine-mapping analysis. We examined the impact of imputation accuracy (measured by AR^2) on association *P*-values and found that poorly imputed variants tended to have a larger association *P*-value (Fig. 2.5C). This trend reduced the chance of getting false positives from low-quality imputation and provided additional support for the dominance association signal at *RUNX2* with milk yield.

Genomic prediction incorporating dominance and imprinting effects

We compared prediction performance of three models: 1) additive effect only (ADD), 2) additive and dominance effects (ADD+DOM), and 3) additive, dominance, and imprinting effects (ADD+DOM+IMP). Overall, the three models showed similar prediction accuracy and unbiasedness for all the eight traits (Fig. 2.7), even though non-additive effects explained >30% of total genetic variance for the three reproduction traits (DPR, CCR, and HCR). A small increase of prediction accuracy for three production traits (<1%) was observed with the models ADD+DOM and ADD+DOM+IMP compared to the model ADD. Paired t-tests showed that the increases were significant (*P*<0.05). However, there was no significant difference in prediction accuracy between the models ADD+DOM and ADD+DOM+IMP for the three traits.

Discussion

This study provided a systematic view of dominance and imprinting effects through a comprehensive analysis of a large cattle data set, including variance decomposition, GWAS, and genomic prediction. The study of imprinting effects benefited from the large size of the cattle data which included complete pedigree, representing one of the largest pedigrees available in a mammalian species, to infer parent-of-origin of alleles. The current study provided another demonstration of the power of dairy industry-oriented data to facilitate biological research (Decker, 2015; Ma *et al*, 2015).

In general, our results are consistent with previous studies regarding the proportion of phenotypic variance explained by dominance effects for complex traits in cattle (Sun *et al*, 2014) and the low heritability of reproduction traits (Liu *et al*, 2008). The U.S. national evaluation includes a regression on inbreeding to account for the effect of dominance on the mean, not just the variance and covariance. Sun et al (2014) found a large advantage in predicting progeny performance by multiplying this regression on inbreeding by estimated genomic inbreeding of the calf, but found only small additional advantage by including dominance variance matrix. However, imprinting effects have been rarely evaluated in livestock studies, and our analysis provided useful information on the contribution of imprinting effects to dairy traits. First, despite their small proportion relative to the total variance, imprinting effects had a significant, non-zero contribution to the phenotypic variation for most of the traits investigated, including all the three production traits and three reproduction traits. Second, imprinting effects explained a much larger proportion of the total genetic variance for reproduction traits than for production traits. These results raised two important questions: does imprinting universally contribute to complex traits, and why are reproduction traits more affected by imprinting? It is worth mentioning that the reproduction traits considered here model pregnancy as a trait of the dam, whereas pregnancy as a trait of the embryo might have a stronger connection to dominance and imprinting.

In this study, we didn’t observe much improvement of prediction accuracy by including dominance and imprinting effects in genomic selection models. This observation can be attributable to a few things: 1) low heritability of non-additive effects; and 2) lacking of full-sib pairs between reference and prediction populations because full-sibs are the primary source of non-additive relationships but dairy data consist of mostly half-sibs.

Using a GWAS approach, we found a dominance association signal and validated it in independent samples. The fine-mapping analysis further confirmed the dominance QTL to be near *RUNX2*, but it was difficult to distinguish causal variants from linked markers. Due to a very small effective population size and a limited number of haplotypes in the dairy cattle population, our imputation works well, even from 50k or less SNP data to sequence-level variants, in our fine-mapping association analysis.

Our study demonstrated the possibility of identifying non-additive effects in GWAS using a large dataset. Additionally, the power of the two-step GWAS approach was comparable to a full mixed-model based method (Table 2.1). The two-step method used in this study was an efficient alternative to identify non-additive effects when fast implementations of full mixed-models are not available. For genomic prediction, we observed a very small but significant increase of prediction accuracy for production traits, but no difference for reproduction traits, when non-additive effects were included. Due to possible sparseness of dominance and imprinting effects, GREML may underperform for prediction and Bayesian models assuming a few large QTLs may perform better. Future studies are needed to develop more accurate prediction models for non-additive effects.

Conclusions

In this study, we comprehensively evaluated the contribution of dominance and imprinting effects to complex traits in dairy cattle. We reported significant, non-zero contributions from dominance and imprinting effects for both production and reproduction traits. The imprinting effects contribute a larger proportion to reproduction traits that production traits. Using GWAS, we identified and validated a dominance association signal with milk yield near *RUNX2*. However, we observed minor increases in prediction accuracy when including non-additive effects in the genomic selection models.

Methods

Genotype and phenotype data

The large dairy cattle database maintained by CDCB and USDA-AGIL includes more than one million genotyped animals with complete pedigree. The data were collected on a continuous basis, and this study included all the Holstein data available until September, 2015. From the database, we extracted 262,757 genotyped females whose sire and dam were also genotyped. The genotypes were generated from 16 different SNP arrays with SNP number ranging from 7K to 50K. The SNP genotypes of all 262,757 females were phased to determine the parent-of-origin of each allele. We first used parent genotypes to phase a SNP genotype of a cow (Ma *et al*, 2015). If this step failed, we then applied a population-based phasing approach using FindHap version 3.0 (VanRaden *et al*, 2013). After phasing, all individuals were imputed to 50K SNP data. When building genomic relationship matrices (GRMs), we further filled a small portion of genotypes that were still missing after imputation from FindHap by randomly sampling genotypes from a multinomial distribution with probabilities of the three genotypes derived under an assumption of Hardy–Weinberg equilibrium.

Among the 262,757 Holstein cows, 42,701 of them had yield deviation (YD) phenotypic data. YD phenotypes were adjusted for appropriate covariates, including farm, year, and season effects. Eight traits were analyzed, including milk yield (MY), fat yield (FY), protein yield (PY), somatic cell score (SCS; a measure of mammary gland health), standardized productive life (STPL; a measure of longevity), daughter pregnancy rate (DPR; a measure of fertility), cow conception rate (CCR; a measure of fertility), and heifer conception rate (HCR; a measure of fertility). Since many cows were not measured for all the phenotypes, the final sample size for the eight traits ranged from 12,911 (STPL) to 29,811 (MY, FY and PY), as shown in Table 2.1.

Variance decomposition with additive, dominance, and imprinting components

Genetic effects of SNPs can be decomposed into three components (i.e., genotypic additive, dominance, and imprinting values), following their evident biological meanings:

$$\begin{bmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{bmatrix} = R + a\begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \end{bmatrix} + d\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} + i\begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} = R + A + D + I \,, \tag{2.1}$$

where G_{12} is the genetic value for the genotype 12 with a paternal allele 1 and a maternal allele 2 (similar for G_{11}, G_{21} and G_{22}), R is the overall mean, a is additive effect, d is dominance effect, i is imprinting effect, A is genotypic additive value arising from a, D is genotypic dominance value arising from d, and I is genotypic imprinting value arising from i. Under Hardy-Weinberg equilibrium, equation (2.1) can be further centralized regarding a and d into

$$\begin{bmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{bmatrix} = R^{*} + a\begin{bmatrix} -2p \\ q-p \\ q-p \\ 2q \end{bmatrix} + d\begin{bmatrix} -2pq \\ 1-2pq \\ 1-2pq \\ -2pq \end{bmatrix} + i\begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} = R^{*} + A^{*} + D^{*} + I \,, \tag{2.2}$$

where R^{*} is the overall mean after centering, p is the frequency of allele 2 and q is the frequency of allele 1, and A^{*} (D^{*}) is genotypic additive (dominance) value after centralization. Note that in equation (2.2), genotypic additive value (A^{*}) is not independent of genotypic dominance value (D^*), or $Cov(A^*, D^*) \neq 0$. To address the issue, we can use the extended natural and orthogonal interactions (NOIA) model (Álvarez-Castro, 2015) under Hardy-Weinberg equilibrium,

$$\begin{bmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{bmatrix} = R^{**} + \beta \begin{bmatrix} -2p \\ q-p \\ q-p \\ 2q \end{bmatrix} + d \begin{bmatrix} -2p^2 \\ 2pq \\ 2pq \\ -2q^2 \end{bmatrix} + i \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} = R^{**} + A^{**} + D^{**} + I \,, \qquad (2.3)$$

where R^{**} is the overall mean and β is allele substitution effect. Despite its similarity to equation (2.2), equation (2.3) results in different variance decomposition. The three components for β, d, and i correspond to breeding value (A^{**}), dominance deviation (D^{**}), and genotypic imprinting value (I), respectively.

The differences and relationships between equations (2.2) and (2.3) have been thoroughly discussed in a previous study (Vitezica *et al*, 2013), although that study did not include imprinting effects. The equation still holds when imprinting effects are included because the genotypic imprinting value is independent of the other two components in both equations (2.2) and (2.3). In theory, the sum of individual breeding value and dominance deviation in equation (2.3) is equal to the sum of individual genotypic additive and dominance values in equation (2.2); and when ignoring the covariance between additive and dominance effects, the sum of additive and dominance genetic variances resulting from the decomposition by equation (2.3) is equal to the sum of the variances of genotypic additive and dominance values resulting from the decomposition by equation (2.2). Additionally, individual genotypic imprinting value in equation (2.2) is the same as in equation (2.3), thus asserting the equivalence of imprinting variance components in the two equations.

The theory holds for multiple loci when assuming linkage equilibrium and independent marker effects (Vitezica *et al*, 2013).

Although it is possible to directly fit SNP effects in a model (Zhu *et al*, 2015b), fitting individual-level genetic components is more efficient, especially for a large dataset with many SNP markers. In this study, we used the following model

$$\begin{gathered}\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{a} + \mathbf{d} + \mathbf{i} + \mathbf{e} \\ \text{with } \mathbf{a} \sim N(0, \sigma_a^2\mathbf{A}),\ \mathbf{d} \sim N(0, \sigma_d^2\mathbf{D}),\ \mathbf{i} \sim N(0, \sigma_p^2\mathbf{P}),\ \mathbf{e} \sim N(0, \sigma_e^2\mathbf{I})\end{gathered} , \qquad (2.4)$$

where the phenotypic value of individuals (***y***) was decomposed into fixed effects (***b***), genotypic additive value (***a***), genotypic dominance value (***d***), genotypic imprinting value (***i***), and residual (***e***). Equation (2.4) can be readily solved by a multi-component restricted maximum likelihood (REML) approach as implemented in GCTA (Yang *et al*, 2011a), as long as we know the covariance structures of the three components, **A**, **D**, and **P**. Different forms of additive genomic relationship matrix (GRM) have been proposed. We used a version with pooled variance across all markers (VanRaden, 2008),

$$A_{ij} = \sum\nolimits_k (Z_{ik} - 2p_k)(Z_{jk} - 2p_k) \Big/ \sum\nolimits_k 2p_k(1-p_k) \qquad (2.5)$$

where Z_{ik} (Z_{jk}) is the additive genotype code for marker k of individual i (j) as shown in the vector corresponding to a in equation (2.1) and p_k is the population frequency of allele 2. Similarly, based on the equivalence of SNP-BLUP and GBLUP (Da *et al*, 2014; Stranden and Garrick, 2009), we can obtain corresponding GRMs for dominance (**D**) and imprinting (**P**), which are shown as following:

$$D_{ij} = \frac{\sum_k \left[H_{ik} - 2p_k(1-p_k)\right]\left[H_{jk} - 2p_k(1-p_k)\right]}{\sum_k 2p_k(1-p_k)\left[1 - 2p_k(1-p_k)\right]} \qquad (2.6)$$

$$P_{ij} = \sum\nolimits_k S_{ik} S_{jk} \Big/ \sum\nolimits_k 2 p_k (1 - p_k) \qquad (2.7)$$

where *H* and *S* are the genotype codes for dominance and imprinting effects as shown in the corresponding vectors in equation (2.1), respectively. Equation (2.6) has been used in previous studies (Su *et al*, 2012; Sun *et al*, 2014). When building GRMs, we used whole-genome markers with minor allele frequency (MAF) ≥0.01. Finally, the software MMAP (O'Connell, 2015), which efficiently implements REML, was used to fit model (2.4).

For comparison purposes, we also performed variance decomposition based on equation (2.3). In this case, we need to use a different dominance GRM ($\mathbf{D}^*$),

$$D_{ij}^* = \sum\nolimits_k H_{ik}^* H_{jk}^* \Big/ \sum\nolimits_k \left(2 p_k (1 - p_k) \right)^2 , \qquad (2.8)$$

where *H** is the dominance genotype code as shown in the vector corresponding to *d* in equation (2.3). Accordingly, the total genetic variance is decomposed to classical additive and dominance genetic variances and variance of genotypic imprinting effect. We further compared the two different kinds of variance decompositions regarding estimates of individual effects and variance components to verify the theory on their equivalence of explaining phenotypes.

Simulation study for validating variance decomposition

Note that when building the GRMs, we assumed that the traits are highly polygenic for the additive, dominance, and imprinting effects. Although the polygenic architecture of additive effects is commonly used for complex traits (Kemper and Goddard, 2012), we have less knowledge on whether dominance and imprinting effects are also polygenic. To address this issue, we simulated a number of datasets to investigate whether model (2.4) can capture dominance and imprinting effects when there are a small number of corresponding QTLs. Specifically, we first obtained a random subsample of 10,000 from the 42,000 cows being analyzed, and then randomly selected markers from the 50k SNPs as additive, dominance, or imprinting QTLs. We simulated QTL effects using a normal distribution and added them up to obtain *a*, *d*, and *i* for each of the 10,000 cows. Thereafter we calculated $\sigma_a^2 = \mathrm{var}(a)$, $\sigma_d^2 = \mathrm{var}(d)$, and $\sigma_p^2 = \mathrm{var}(i)$ using corresponding simulated genetic values. Based on the heritability we set to simulate, we calculated σ_e^2 and simulated *e* by sampling it from $N(0, \sigma_e^2)$. The phenotype for each individual animal was simulated by adding up *a*, *d*, *i*, and *e*.

To ensure realistic simulations, we picked variance of the normal distribution for simulating effect sizes so the variance decomposition was the same between simulated and real data. Our simulation scenarios included two representative traits, milk yield and DPR, separately. Three scenarios were simulated for either trait by varying QTL numbers, including 1000+10+10 (1000, 10 and 10 QTLs for additive, dominance, and imprinting effects, respectively), 1000+100+100, and 1000+1000+1000. Simulation for each scenario was repeated 100 times. We fitted model (2.4) for each simulated data set and compared variance component estimation between the three scenarios.

Genome-wide association study of non-additive effects

To increase computational efficiency, we used a two-step strategy for genome-wide association study, similar to the GRAMMAR approach (Aulchenko *et al*, 2007). First, we fitted model (2.4), and obtained the residuals to adjust for polygenic effects. Second, we used the residuals as response variable to fit a multiple linear regression model for each SNP,

$$\mathbf{e} = \mu + a_k\mathbf{Z}_k + d_k\mathbf{H}_k + i_k\mathbf{S}_k + \boldsymbol{\varepsilon}, \tag{2.9}$$

where $\mathbf{Z}_k$, $\mathbf{H}_k$ and $\mathbf{S}_k$ are the genotype codes of marker k for additive, dominance and imprinting effects, respectively, as described in equations (2.5, 2.6, 2.7), and a_k, d_k, and i_k are corresponding SNP effects. SNPs were filtered by MAF ≥0.01 and P-value of Chi-square test for Hardy–Weinberg equilibrium $\geq 1 \times 10^{-6}$. Association P-values were calculated from t-tests for the three types of SNP effects.

For association signals with sufficient statistical evidence from the two-step analysis, we further used the full mixed model,

$$\begin{gathered}\mathbf{y} = \mu + a_k\mathbf{Z}_k + d_k\mathbf{H}_k + i_k\mathbf{S}_k + \mathbf{a} + \mathbf{d} + \mathbf{i} + \mathbf{e} \\ \text{with } \mathbf{a} \sim N(0,\sigma_a^2\mathbf{A}),\ \mathbf{d} \sim N(0,\sigma_d^2\mathbf{D}),\ \mathbf{i} \sim N(0,\sigma_p^2\mathbf{P}),\ \mathbf{e} \sim N(0,\sigma_e^2\mathbf{I})\end{gathered}, \tag{2.10}$$

or its reduced version,

$$\mathbf{y} = \mu + a_k\mathbf{Z}_k + d_k\mathbf{H}_k + i_k\mathbf{S}_k + \mathbf{a} + \mathbf{e} \text{ with } \mathbf{a} \sim N(0,\sigma_a^2\mathbf{A}) \text{ and } \mathbf{e} \sim N(0,\sigma_e^2\mathbf{I}), \tag{2.11}$$

to rerun the association analysis, depending on whether the additive effects can explain a majority of total genetic variance on the trait being analyzed. Here, the response variables in equation (2.10) and (2.11) are yield deviations. Again, we applied the software MMAP (O'Connell, 2015) to fit the mixed models.

Validation of non-additive association signals using independent data

Our discovery GWAS used the data available until September, 2015. From then to April, 2016, we assembled a new dataset to validate the signal found in the initial GWAS. The validation data consisted of 5,514 cows with both genotypes and milk phenotypes. The genotypes in the validation data were phased with the same procedures as used for the discovery data set. With the validation data, model (2.11) was used to analyze associations between milk and 50 SNP markers around the *RUNX2* signal. The GRM was built using all chip SNPs except those on chromosome 23, which resulted in a leave-one-chromosome-out analysis (LOCO) (Yang *et al*, 2014). We also built the GRM using all genome-wide SNPs and compared it with the LOCO analysis. The validation data were also used to analyze the significant dominance associations around Chr5:107,000,000 with both fat and protein. The three SNPs with the smallest discovery *P*-value were analyzed with model (2.11) for fat and protein, respectively.

Fine mapping for the *RUNX2* dominance signal

First, we used the sequence data of 443 Holstein bulls from the 1000 Bull Genomes project (Daetwyler *et al*, 2014) (Run 5.0) to check LD levels between the targeted SNP (Chr23:18,600,456) and SNPs/ biallelic indels around it. Based on the LD decay pattern, we chose the region of ±500 kb from the targeted SNP for fine mapping. Then, we used the sequence genotypes of the 443 bulls as reference to impute the 50k genotypes of 29,811 cows to sequence genotypes. Beagle version 4 (Browning and Browning, 2013) was used for the imputation with default parameters. To increase accuracy, our imputation covered a larger region of ±1 Mb from the targeted SNP. After imputation, we removed non-informative SNPs, i.e. SNPs with a MAF <0.01, SNPs with a *P*-value of Chi-square test for Hardy–Weinberg equilibrium $< 1 \times 10^{-6}$ and SNPs with an allelic R^2 (AR^2) <0.05. AR^2, reported by Beagle software, is the estimated squared correlation between the most probable alternative allele dose and the true alternative allele dose and serves as a good metric for estimating imputation accuracy (Browning and

Browning, 2009). The analysis of associations between milk and the imputed sequence variants within the targeted region (Chr23:18,100,456-19,100,456) was performed with a two-step method as described in our GWAS section.

Genomic Prediction

We estimated the values of the three effects for individuals in the training population from fitting model (2.4) in MMAP. The genomic predictions for new individuals can be calculated by

$$\hat{\mathbf{g}}_n = \hat{\boldsymbol{\alpha}}_n + \hat{\mathbf{d}}_n + \hat{\mathbf{i}}_n = \mathbf{A}_{n\times t}\mathbf{A}^{-1}_{t\times t}\hat{\mathbf{a}}_t + \mathbf{D}_{n\times t}\mathbf{D}^{-1}_{t\times t}\hat{\mathbf{d}}_t + \mathbf{P}_{n\times t}\mathbf{P}^{-1}_{t\times t}\hat{\mathbf{i}}_t \, , \tag{2.12}$$

where the subscripts *n* and *t* indicate the sets of new individuals and training population, respectively. Besides model (2.4) (ADD+DOM+IMP), we also considered two reduced models, the additive model (ADD) and the additive-plus-dominance model (ADD+DOM), and compared the prediction performance between the three models. Ten-fold cross validation was used to assess 1) prediction accuracy, defined as the Person correlation between genomic estimated breeding value (GEBV) and phenotype, and 2) unbiasedness, defined as the regression coefficient of phenotype on GEBV in the validation population.

References

Aliloo H, Pryce J, González-Recio O, Cocks B, Goddard M, Hayes B (2016a). Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. *Journal of Dairy Science*.

Aliloo H, Pryce JE, González-Recio O, Cocks BG, Hayes BJ (2016b). Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. *Genetics Selection Evolution* **48**(1)**:** 8.

Álvarez-Castro JM (2015). Dissecting genetic effects with imprinting. *Models and Estimation of Genetic Effects***:** 35.

Aulchenko YS, de Koning DJ, Haley C (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**(1)**:** 577-585.

Browning BL, Browning SR (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**(2)**:** 210-223.

Browning BL, Browning SR (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**(2)**:** 459-471.

Carlborg Ö, Haley CS (2004). Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics* **5**(8)**:** 618-625.

Da Y, Wang C, Wang S, Hu G (2014). Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS One* **9**(1)**:** e87666.

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF *et al* (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46**(8)**:** 858-865.

Decker JE (2015). Agricultural Genomics: Commercial Applications Bring Increased Basic Research Power. *PLoS Genet* **11**(11)**:** e1005621.

Guo X, Christensen OF, Ostersen T, Wang Y, Lund MS, Su G (2016). Genomic prediction using models with dominance and imprinting effects for backfat thickness and average daily gain in Danish Duroc pigs. *Genetics Selection Evolution* **48**(1)**:** 67.

Hill WG, Goddard ME, Visscher PM (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* **4**(2)**:** e1000008.

Hu Y, Rosa GJ, Gianola D (2015). A GWAS assessment of the contribution of genomic imprinting to the variation of body mass index in mice. *BMC genomics* **16**(1)**:** 576.

Kemper KE, Goddard ME (2012). Understanding and predicting complex traits: knowledge from cattle. *Hum Mol Genet* **21**(R1)**:** R45-51.

Liu Z, Jaitner J, Reinhardt F, Pasman E, Rensing S, Reents R (2008). Genetic evaluation of fertility traits of dairy cattle using a multiple-trait animal model. *J Dairy Sci* **91**(11)**:** 4333-4343.

Lopes MS, Bastiaansen JW, Janss L, Knol EF, Bovenhuis H (2015). Estimation of additive, dominance, and imprinting genetic variance using genomic data. *G3: Genes| Genomes| Genetics* **5**(12)**:** 2629-2637.

Ma L, Brautbar A, Boerwinkle E, Sing CF, Clark AG, Keinan A (2012). Knowledge-Driven Analysis Identifies a Gene-Gene Interaction Affecting High-Density Lipoprotein Cholesterol Levels in Multi-Ethnic Populations. *Plos Genet* **8**(5).

Ma L, O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C *et al* (2015). Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLoS genetics* **11**(11)**:** e1005387.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ *et al* (2009). Finding the missing heritability of complex diseases. *Nature* **461**(7265)**:** 747-753.

Moore T, Haig D (1991). Genomic imprinting in mammalian development: a parental tug-of-war. *Trends in Genetics* **7**(2)**:** 45-49.

Nishio M, Satoh M (2015). Genomic best linear unbiased prediction method including imprinting effects for genomic evaluation. *Genetics Selection Evolution* **47**(1)**:** 32.

O’Connell JR (2015). MMAP User Guide. Available: http://edn.som.umaryland.edu/mmap/index.php. Accessed 8 October 2015.

Owens TW, Rogers RL, Best SA, Ledger A, Mooney AM, Ferguson A *et al* (2014). Runx2 is a novel regulator of mammary epithelial cell fate in development and breast cancer. *Cancer research* **74**(18)**:** 5277-5286.

Stranden I, Garrick DJ (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J Dairy Sci* **92**(6)**:** 2971-2975.

Su G, Christensen OF, Ostersen T, Henryon M, Lund MS (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS One* **7**(9)**:** e45293.

Sun C, VanRaden P, O’Connell J, Weigel K, Gianola D (2013). Mating programs including genomic relationships and dominance effects. *Journal of dairy science* **96**(12)**:** 8014-8023.

Sun C, VanRaden PM, Cole JB, O'Connell JR (2014). Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. *PLoS One* **9**(8)**:** e103934.

VanRaden P (2016). Practical implications for genetic modeling in the genomics era. *Journal of dairy science* **99**(3)**:** 2405-2412.

VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**(11)**:** 4414-4423.

VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB *et al* (2013). Genomic imputation and evaluation using high-density Holstein genotypes. *J Dairy Sci* **96**(1)**:** 668-678.

Vitezica ZG, Varona L, Legarra A (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* **195**(4)**:** 1223-1230.

Wang C, Prakapenka D, Wang S, Pulugurta S, Runesha HB, Da Y (2014). GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. *BMC bioinformatics* **15**(1)**:** 270.

Wittenburg D, Melzer N, Reinsch N (2015). Genomic additive and dominance variance of milk performance traits. *Journal of Animal Breeding and Genetics* **132**(1)**:** 3-8.

Xiang T, Christensen OF, Vitezica ZG, Legarra A (2016). Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. *Genetics Selection Evolution* **48**(1)**:** 92.

Yang J, Lee SH, Goddard ME, Visscher PM (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**(1)**:** 76-82.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**(2)**:** 100-106.

Zhu Z, Bakshi A, Vinkhuyzen AA, Hemani G, Lee SH, Nolte IM *et al* (2015). Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet* **96**(3)**:** 377-385.

Tables

Table 2.1. Variance decomposition of genotypic additive, dominance, and imprinting values for eight dairy traits.

Trait	*N*	Proportion in Phenotypic Variance (SE)				Proportion in Total Genetic Variance			*P*-value of test for σ^2=0		
		A	D	I	H^2	A	D	I	A	D	I
MY	29811	0.338 (0.009)	0.040 (0.005)	0.008 (0.002)	0.386 (0.009)	0.875	0.104	0.020	3.5×10^{-151}	9.9×10^{-15}	4.9×10^{-4}
FY	29811	0.312 (0.009)	0.025 (0.005)	0.004 (0.002)	0.340 (0.009)	0.917	0.073	0.010	3.9×10^{-145}	1.1×10^{-7}	0.04
PY	29811	0.272 (0.009)	0.040 (0.005)	0.007 (0.002)	0.319 (0.009)	0.853	0.126	0.021	1.8×10^{-122}	1.3×10^{-13}	2.5×10^{-3}
SCS	29392	0.102 (0.007)	0.010 (0.006)	0.002 (0.002)	0.114 (0.007)	0.893	0.087	0.019	2.2×10^{-48}	0.04	0.14
STPL	12911	0.031 (0.007)	0.000 (0.011)	0.000 (0.004)	0.031 (0.010)	1.0	0.0	0.0	3.4×10^{-06}	0.5	0.5
DPR	22942	0.044 (0.006)	0.011 (0.007)	0.015 (0.004)	0.069 (0.008)	0.637	0.154	0.209	5.2×10^{-15}	0.07	1.9×10^{-5}
CCR	14318	0.051 (0.008)	0.007 (0.011)	0.021 (0.005)	0.079 (0.011)	0.647	0.090	0.264	2.2×10^{-11}	0.27	6.0×10^{-5}
HCR	28601	0.008 (0.003)	0.002 (0.005)	0.005 (0.002)	0.014 (0.005)	0.538	0.108	0.354	3.5×10^{-3}	0.39	0.01

MY: milk yield. FY: fat yield; PY: protein yield. SCS: somatic cell score. STPL: standardized productive life. DPR: daughter pregnancy rate. CCR: cow conception rate. HCR: heifer conception rate. *N*: sample size. A: additive effect. D: dominance effect. I: imprinting effect. SE: standard error. H^2: broad-sense heritability.

Table 2.2. Variance decomposition of breeding value, dominance deviation and genotypic imprinting value for eight dairy traits

Trait	N	Proportion in Phenotypic Variance (SE)				Proportion in Total Genetic Variance			P-value of test for σ^2=0		
		A**	D**	I**	H^2	A**	D**	I**	A**	D**	I**
MY	29811	0.348 (0.009)	0.030 (0.004)	0.008 (0.002)	0.386 (0.009)	0.902 (0.011)	0.078 (0.010)	0.021 (0.006)	3.2E-163	8.2E-15	4.4E-04
FY	29811	0.318 (0.009)	0.019 (0.004)	0.004 (0.002)	0.341 (0.009)	0.934 (0.012)	0.056 (0.010)	0.010 (0.006)	1.9E-153	3.5E-08	4.1E-02
PY	29811	0.281 (0.009)	0.031 (0.004)	0.007 (0.002)	0.320 (0.009)	0.880 (0.014)	0.098 (0.012)	0.022 (0.008)	1.0E-133	2.3E-14	2.2E-03
SCS	29392	0.105 (0.006)	0.006 (0.004)	0.002 (0.002)	0.113 (0.008)	0.928 (0.038)	0.053 (0.035)	0.019 (0.018)	2.3E-54	7.5E-02	1.4E-01
STPL	12911	0.031 (0.006)	0.004 (0.009)	0.000 (0.004)	0.034 (0.011)	0.887 (0.261)	0.113 (0.240)	0.000 (0.116)	2.4E-07	3.4E-01	5.0E-01
DPR	22942	0.047 (0.005)	0.006 (0.006)	0.015 (0.004)	0.068 (0.008)	0.695 (0.073)	0.092 (0.076)	0.213 (0.049)	2.6E-19	1.3E-01	1.8E-05
CCR	14318	0.053 (0.007)	0.006 (0.009)	0.021 (0.005)	0.079 (0.012)	0.668 (0.096)	0.070 (0.108)	0.262 (0.067)	3.4E-14	2.7E-01	6.1E-05
HCR	28601	0.008 (0.002)	0.001 (0.004)	0.005 (0.002)	0.014 (0.005)	0.584 (0.223)	0.048 (0.306)	0.368 (0.176)	5.2E-04	4.4E-01	1.1E-02

MY: milk yield. FY: fat yield; PY: protein yield. SCS: somatic cell score. STPL: standardized productive life. DPR: daughter pregnancy rate. CCR: cow conception rate. HCR: heifer conception rate. N: sample size. A**: breeding value. D**: dominance deviation. I**: genotypic imprinting value. SE: standard error. H^2: broad-sense heritability.

Table 2.3. Top two SNPs associated with milk yield near the *RUNX2* gene.

SNP	Chr	Position	MAF	Model	β_A (SE)	*P*-value	β_D (SE)	*P*-value	β_I (SE)	*P*-value
Hapmap48809-BTA-55698	23	17275448	0.15	Two-step	153.4 (33.0)	3.33×10^{-6}	197.1 (37.6)	1.56×10^{-7}	-3.64 (18.0)	0.84
				A	223.6 (51.8)	1.57×10^{-5}	255.2 (44.7)	1.17×10^{-8}	-1.54 (23.8)	0.95
				A+D+I	212.7 (51.6)	3.82×10^{-5}	241.7 (45.3)	9.54×10^{-8}	-0.52 (25.5)	0.98
BovineHD2300004730	23	18600456	0.10	Two-step	207.3 (47.6)	1.31×10^{-5}	273.5 (52.1)	1.54×10^{-7}	10.31 (21.3)	0.63
				A	206.2 (67.6)	2.29×10^{-3}	353.6 (62.3)	1.43×10^{-8}	-3.43 (28.8)	0.91
				A+D+I	200.6 (67.5)	2.96×10^{-3}	340.4 (62.9)	6.33×10^{-8}	7.52 (30.8)	0.81

Chr: chromosome. MAF: minor allele frequency. β: regression coefficient. SE: standard error.

Figures

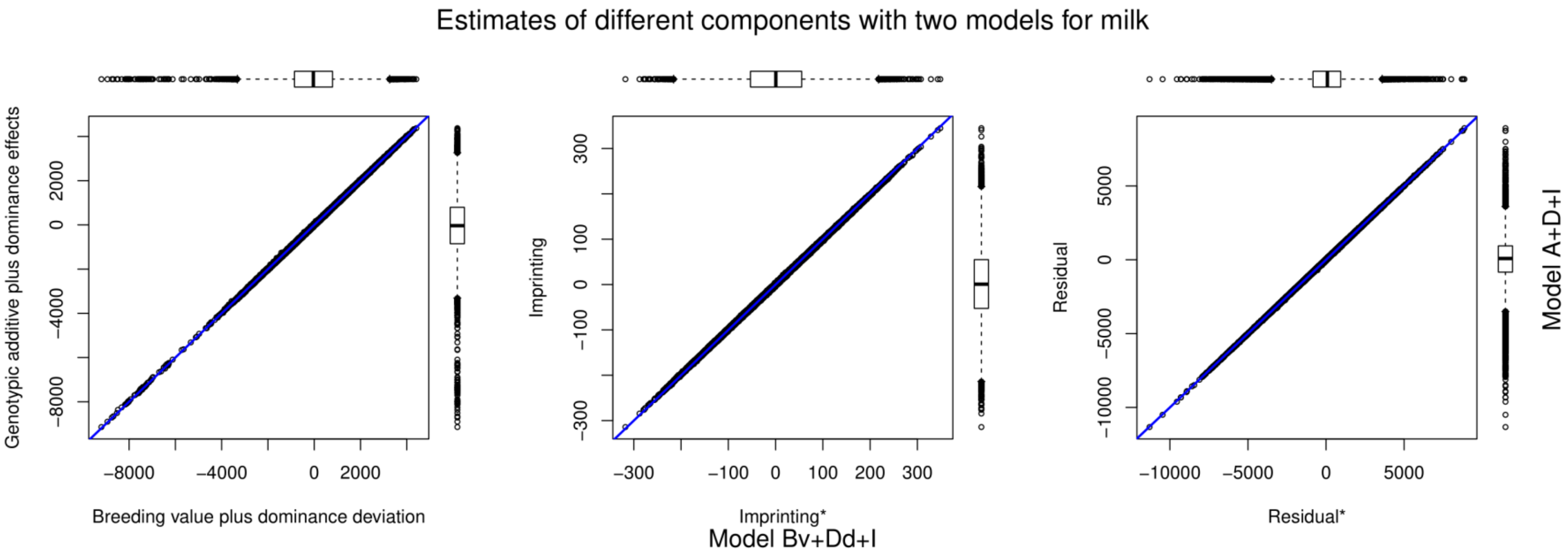


Figure 2.1. Individual estimates of variance components with two decomposition models for milk. Each point indicates the component estimate for each individual. Blue line indicates *y*=*x*. The x-axis shows the components from the model decomposing genetic effect to breeding value, dominance deviation and genotypic imprinting value, while y-axis shows the components from the model decomposing genetic effect to genotypic additive, dominance and imprinting values.

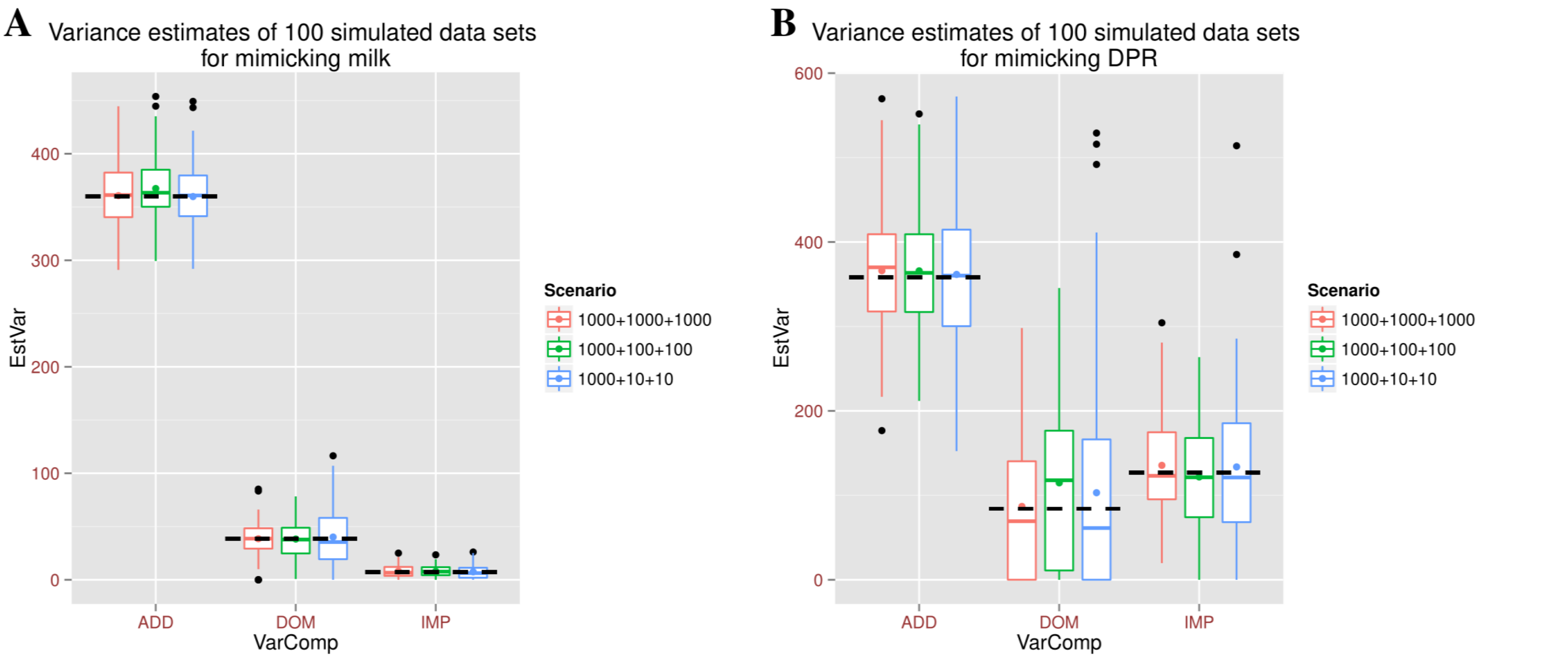


Figure 2.2. Variance decomposition using simulated datasets. The dash line indicates expected value of corresponding variance component.

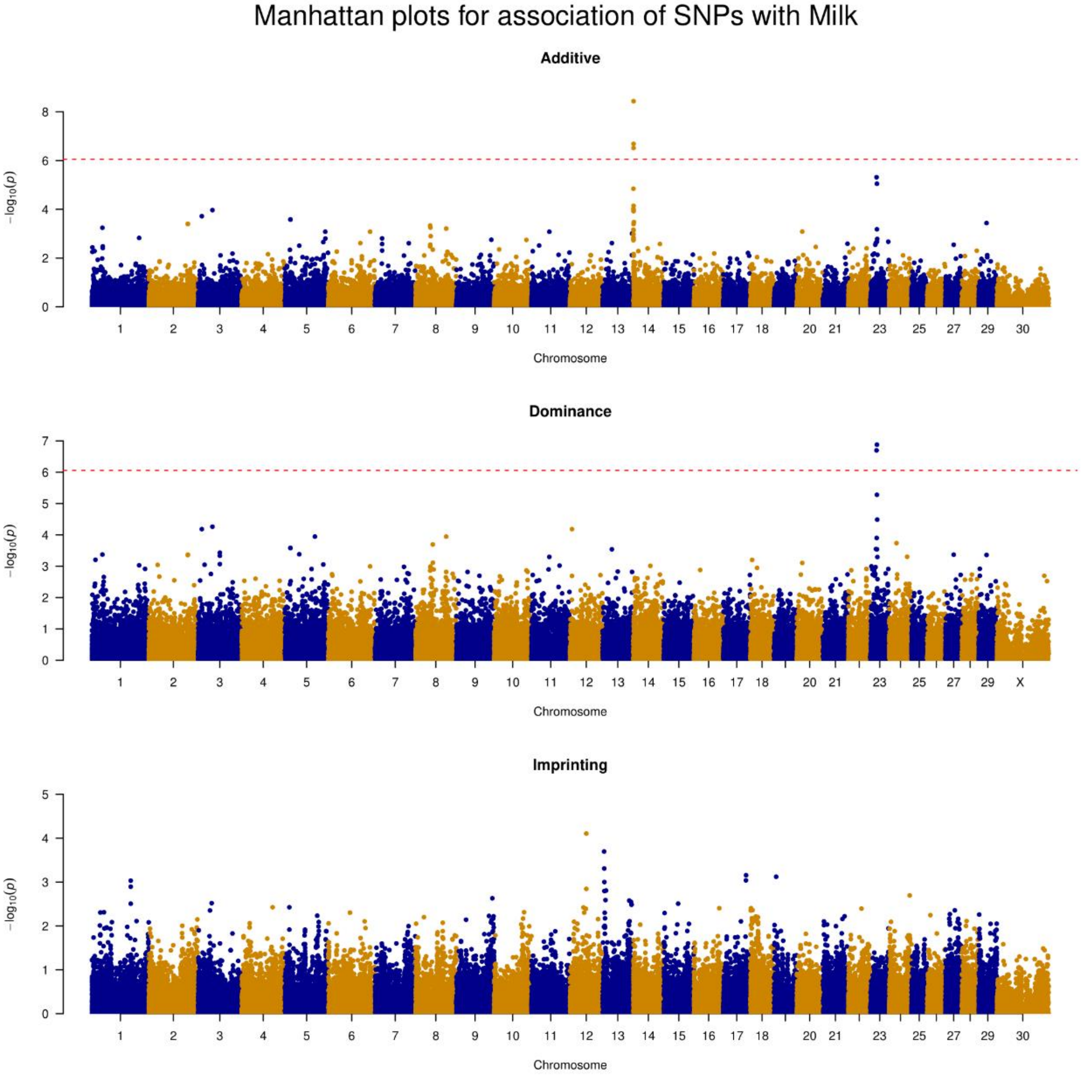


Figure 2.3. Manhattan plots for associations of SNP effects with milk yield.

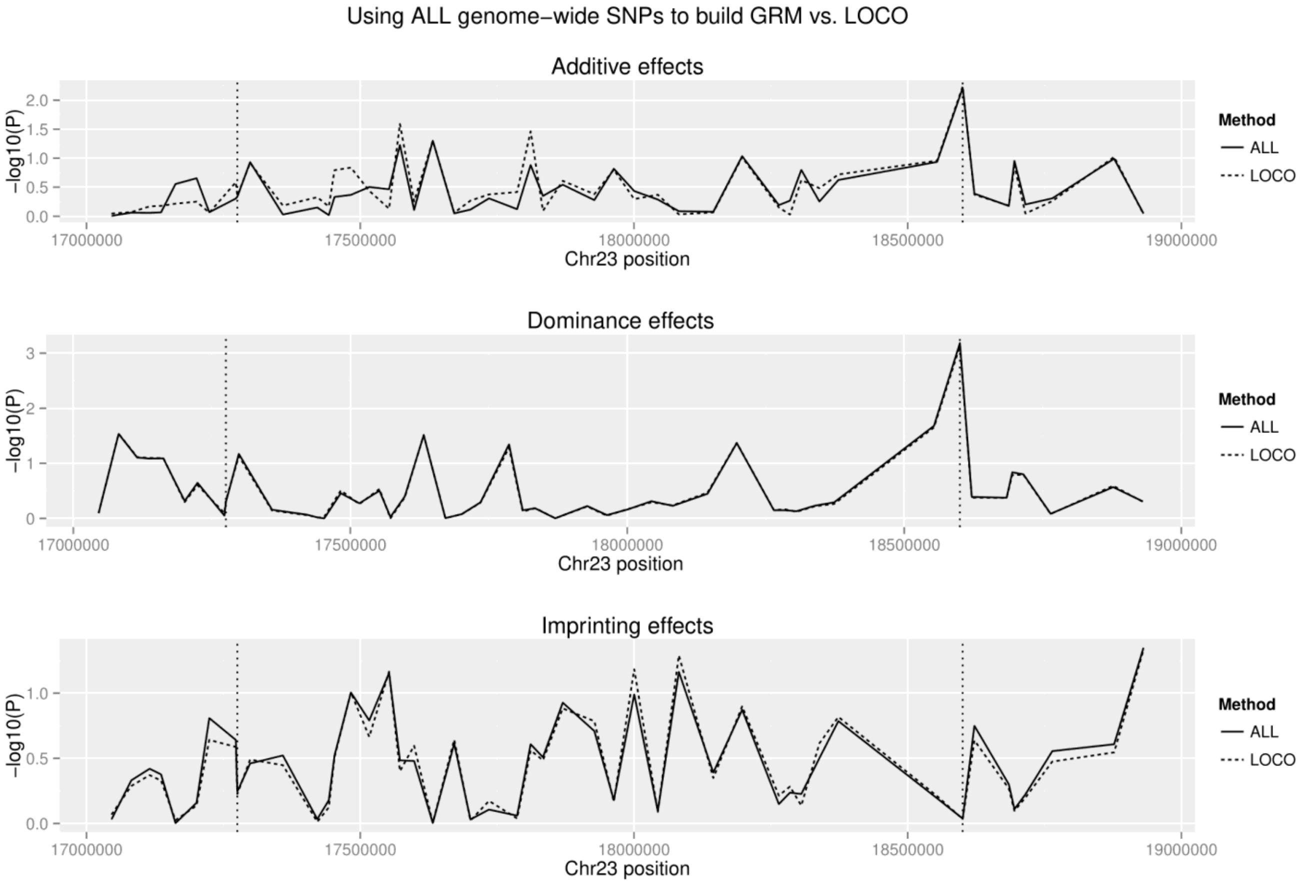


Figure 2.4. Mixed-model based association analysis between milk yield and 50 SNPs around *RUNX2* in the validation data set. The two vertical dash lines indicate SNPs Hapmap48809-BTA-55698 and BovineHD2300004730, respectively.

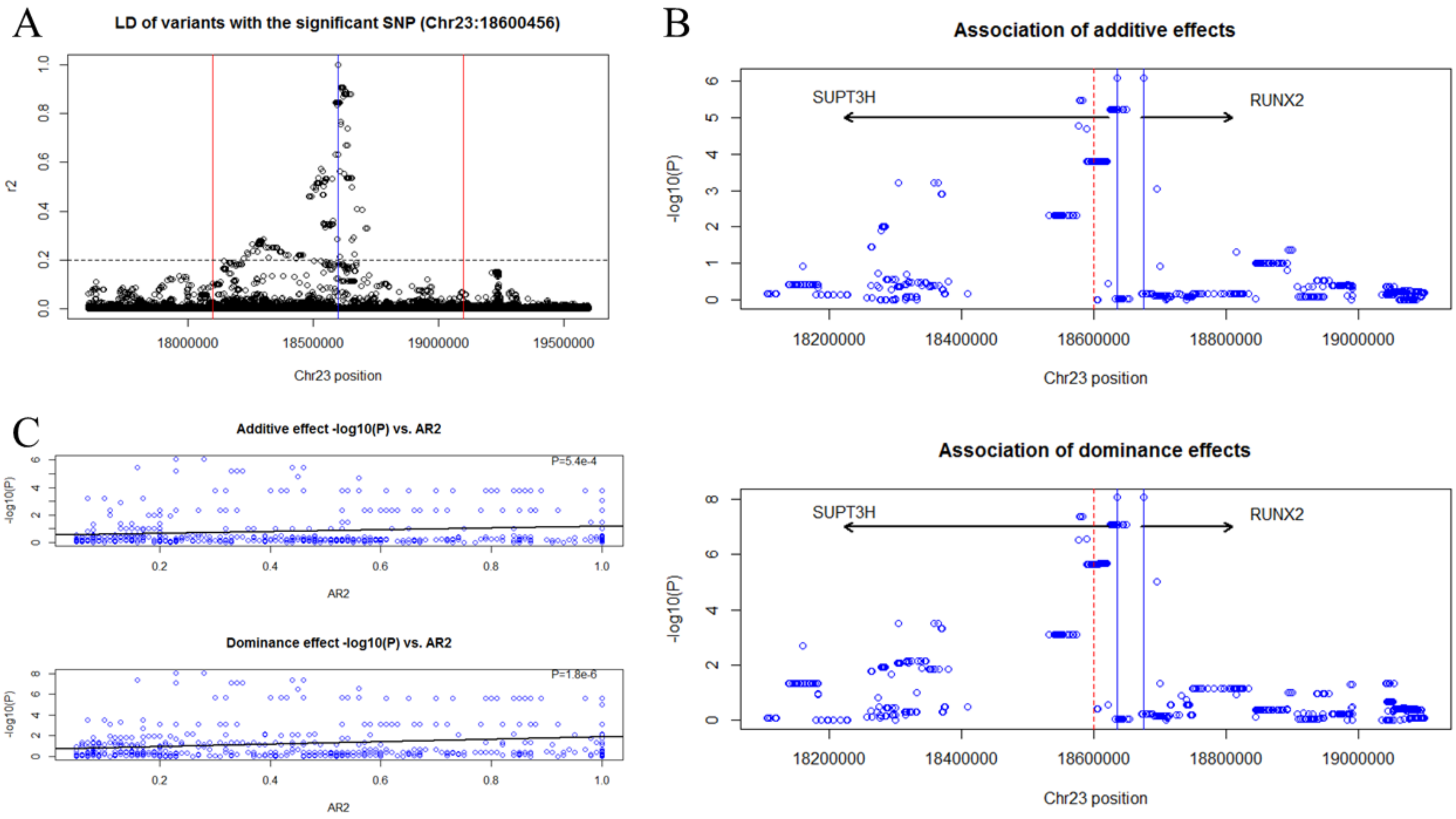


Figure 2.5. Fine-mapping of the dominance association with milk yield near *RUNX2*.

A) LD between BovineHD2300004730 and adjacent variants

B) Association results of additive and dominance effects. The red dash line indicates the target SNP (BovineHD2300004730), while the two blue solid lines indicate the two variants with the smallest *P*-value.

C) The influence of imputation reliability measured by AR^2 on association *P*-values. The black lines indicate the regression line of $-\log_{10}(P)$ on AR^2, and at the right-upper corner are the *P*-values for model fitting of the regression.

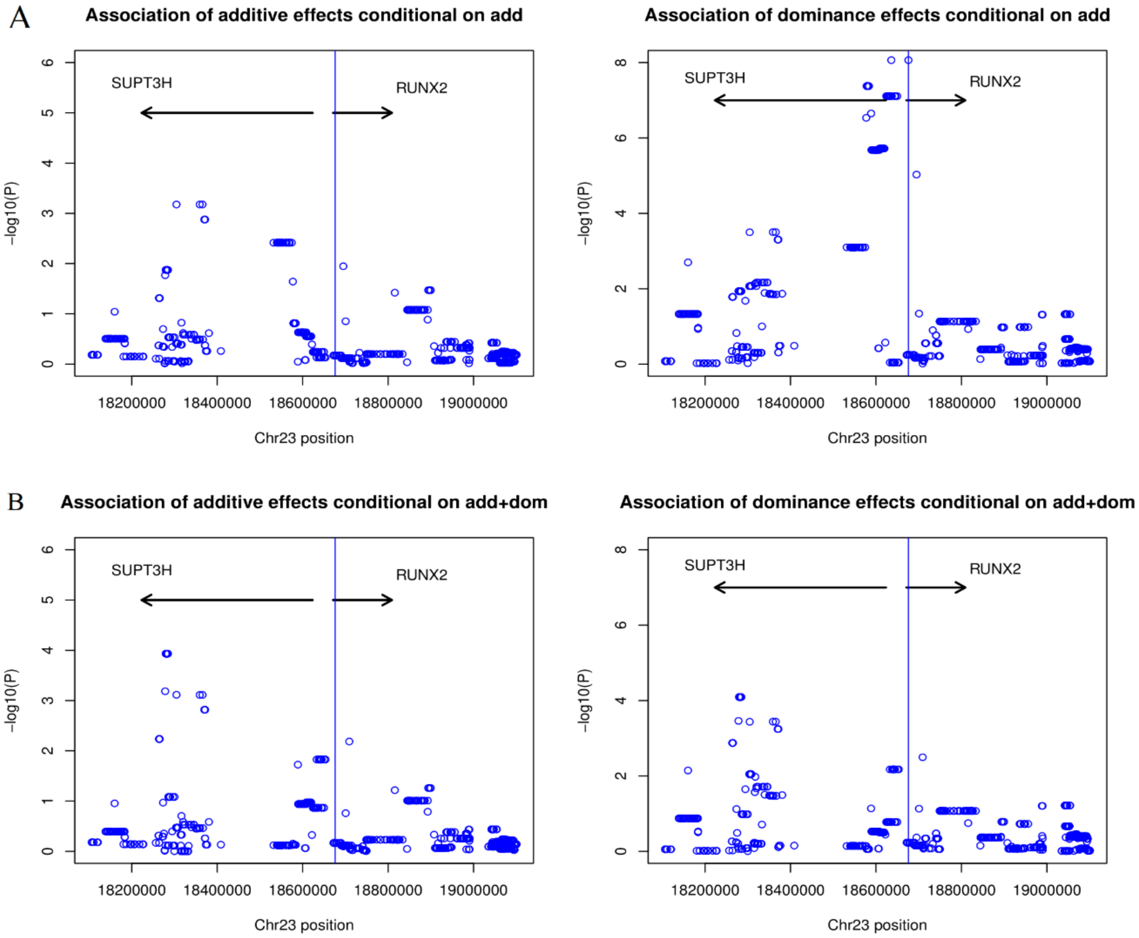


Figure 2.6. Association analysis conditional on the additive effect (A) and both the additive and dominance effects (B) of variant Chr23:18676057. The vertical blue line indicates the location of Chr23:18676057.

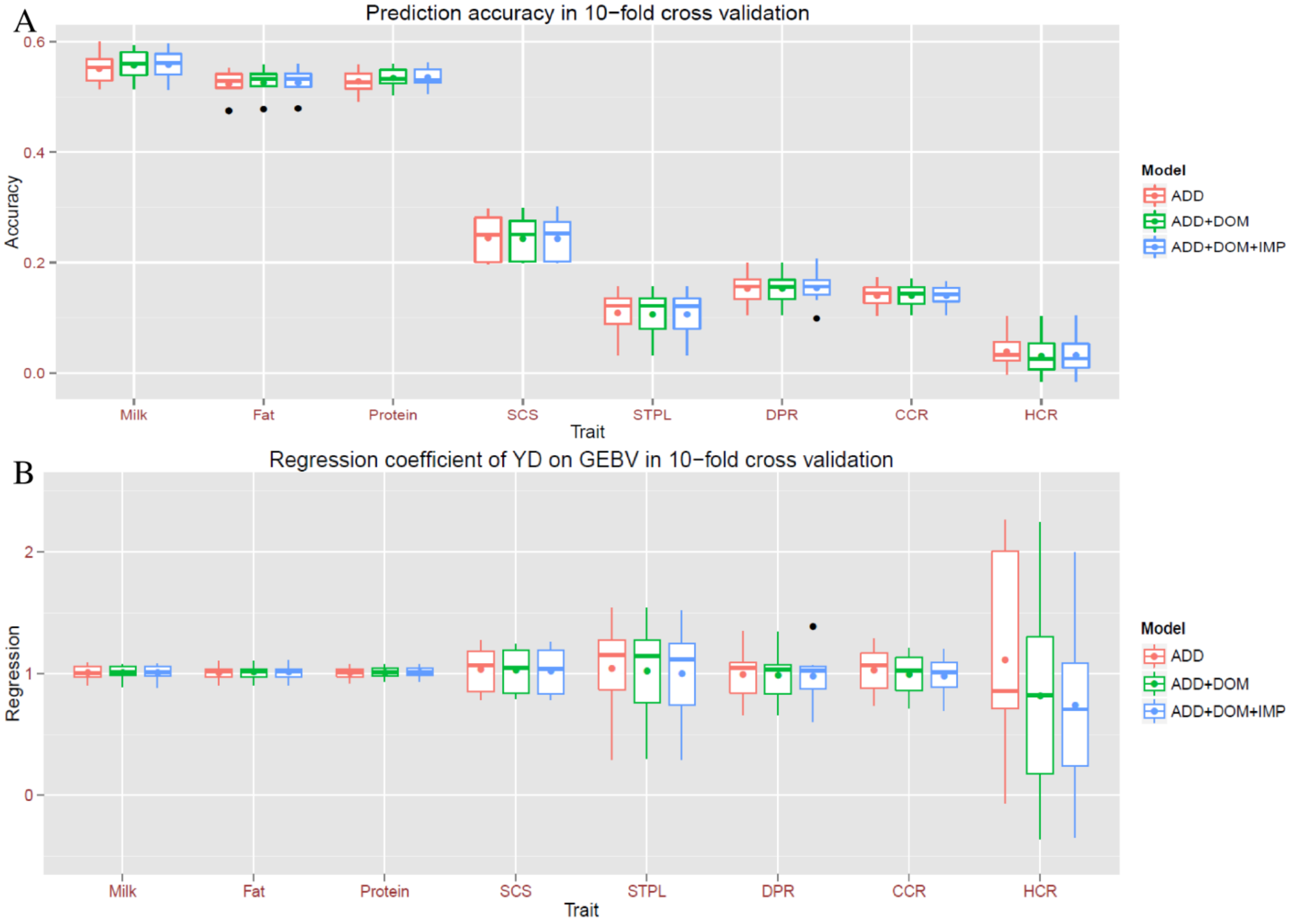


Figure 2.7. Prediction performance of three models for eight dairy traits

Chapter 3: Fast Bayesian Fine-Mapping and SNP-set Association for Population and Pedigree Data

Abstract

Motivation: Routine approaches for genome-wide association studies (GWAS) are generally based on Wald, likelihood ratio, or score tests. Use of Bayes factors is a promising alternative; however, there are currently few fast implementations of such methods for single-marker/SNP-set association analysis. Though Bayesian methods are extensively used in fine-mapping, existing software tools mostly have some drawback, e.g., infeasible model enumeration or insufficient model search.

Results: We propose a unified Bayesian model for single-marker/SNP-set association and fine-mapping and develop a software tool, BFMAP, which can deal with both population and pedigree data. In association tests, it computes not only Bayes factor but also its null distribution, thus also providing *p*-value. In fine-mapping, we implement two fast model search algorithms (forward selection and shotgun stochastic search (SSS)) and introduce simulated annealing to make SSS do more sufficient model search. Furthermore, BFMAP can easily incorporate functional annotation into fine-mapping. We demonstrate that BFMAP achieves a power similar to or higher than existing software tools but is at least a few times faster with respect to single-marker/SNP-set association tests. We also show that BFMAP performs well for fine-mapping even for complex linkage disequilibrium structures.

Introduction

The past decade has witnessed a dramatic advance in our understanding on genetic architecture of complex traits. A variety of computational and statistical approaches have been developed and/or applied for unraveling the genetic cause of phenotypic variations, e.g., genomic-relatedness-based restricted maximum-likelihood (GREML) for estimating SNP heritability (Yang *et al*, 2015; Yang *et al*, 2010; Yang *et al*, 2011b), linear mixed models for genome-wide association studies (GWAS) (Kang *et al*, 2010; Kang *et al*, 2008; Lippert *et al*, 2011; Loh *et al*, 2015b; Svishcheva *et al*, 2012; Yang *et al*, 2014; Zhou and Stephens, 2012), SNP-set kernel association tests (Ionita-Laza *et al*, 2013; Wu *et al*, 2011), Bayesian fine-mapping (Benner *et al*, 2016; Chen *et al*, 2015; Hormozdiari *et al*, 2014; Kichaev *et al*, 2014; Servin and Stephens, 2007), linkage disequilibrium (LD) score regression for estimating SNP heritability and genetic correlation and partitioning heritability (Bulik-Sullivan *et al*, 2015a; Bulik-Sullivan *et al*, 2015b; Finucane *et al*, 2015).

The GWAS approaches, either for single-marker or for SNP-set, are generally based on Wald, likelihood ratio, or score tests. Though use of Bayes factors is a promising alternative (Wakefield, 2009), some problems hinder its application. First, Bayes factor depends on prior, and it is impractical to specify a fixed threshold in all scenarios (in contrast to universal use of *p*-value threshold, 5E-8). Second, it is not easy to specify a proper prior. A diffusive prior tends to favor, unintentionally, the null model, which is so-called Bartlett’s paradox (Bartlett, 1957). To solve the problems, Zhou and Guan (2017) recently derived the null distribution of Bayes factors in linear regression and formulated a novel scaled Bayes factor. However, the linear regression model studies there can only deal with independent samples.

As large-scale sequence data are becoming available, it is now feasible to fine-map a trait to single-variant resolution. Fine-mapping is basically a model selection problem. Many statistical methods have been used to solve the problem, e.g., stepwise selection (Huang *et al*, 2017), exhaustive search limiting maximum model size (Chen *et al*, 2015; Hormozdiari *et al*, 2014; Kichaev *et al*, 2014; Servin and Stephens, 2007), shotgun stochastic search (Benner *et al*, 2016). Stepwise selection is fast and works well for identifying independent causal variants. Exhaustive search is capable of handling all LD structures; however, it is often infeasible when we aim to find multiple causal effects in many variants (e.g. 1000). Shotgun stochastic search (SSS) overcomes this problem by identifying models with high posterior probability and ignoring models with negligible probability (Hans *et al*, 2007). However, SSS may fail to find all important models for some LD structures, even with a long chain. Additionally, most of the existing fine-mapping tools use summary statistics. Though this is a great feature, direct use of genotypes and phenotypes results in exact computation and is more straightforward, especially in some species where summary statistics is not commonly used (e.g. dairy cattle).

To address the aforementioned issues, we propose a unified Bayesian model for single-marker/SNP-set association and fine-mapping which can deal with both population and pedigree data. In this work, we extend the theory of Zhou and Guan (2017) to a Bayesian model which contains a polygenic term to control population structure, making use of Bayes factors in GWAS straightforward. In fine-mapping, we implement forward selection and SSS and introduce simulated annealing to make SSS do more sufficient model search. Furthermore, we develop an approach to incorporate functional annotation into fine-mapping. The approach can be readily applied to many other existing fine-mapping tools. All these methods are implemented in the software tool, BFMAP. We demonstrate that BFMAP achieves a power similar to or higher than existing software tools but is at least a few times faster with respect to single-marker/SNP-set association tests. We also show that BFMAP performs well for fine-mapping even for complex linkage disequilibrium structures.

Methods

Bayesian model

We use a unified Bayesian model for single-marker/SNP-set association and fine-mapping:

$$\begin{gathered}\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{g} + \mathbf{e} \\ \mathbf{b} \sim N(0, \varphi\sigma_e^2\mathbf{I}) \\ \mathbf{a} \sim N(0, \gamma\sigma_e^2\mathbf{A}) \\ \mathbf{g} \sim N(0, \eta\sigma_e^2\mathbf{G}) \\ \mathbf{e} \sim N(0, \sigma_e^2\mathbf{R}) \\ P\left(\sigma_e^2\right) \propto 1/\sigma_e^2\end{gathered}, \tag{3.1}$$

where $\boldsymbol{y}$ is a phenotype vector of size n for a complex trait, $\boldsymbol{b}$ is a vector of covariate (other than genomic variants) effect and $\boldsymbol{X}$ is corresponding design matrix, $\boldsymbol{a}$ is a vector of variant effect with diagonal variance structure $\boldsymbol{A}$ and $\boldsymbol{Z}$ is corresponding genotype coding matrix (e.g., genotype coding for additive, dominance or imprinting effects (Jiang *et al*, 2017)), $\boldsymbol{g}$ is a vector of polygenic effect for controlling population structure and $\boldsymbol{G}$ is corresponding variance structure matrix (e.g., genomic relationship matrix

(GRM)), and $\boldsymbol{e}$ is residual with diagonal variance structure $\boldsymbol{R}$ for modelling reliability or accuracy of phenotypic records. The common variance component (σ_e^2) is given a non-informative Jeffrey's prior. Other variance parameters (φ, γ and η) are treated as known. Generally, we can set φ to a large value (e.g., 1E8) to make $\boldsymbol{b}$ act like fixed effects. A genomic variant is usually considered to be of small but noticeable effect, so we can set γ to 0.01 or 0.04 (Chen *et al*, 2015; Zhou and Guan, 2017b). When $\boldsymbol{Za}$ only accounts for a tiny proportion of phenotypic variance (this is generally true when the variant set of interest is small), we can set η based on heritability (h^2), $\eta = h^2 / 1 - h^2$. In practice, we can instead use heritability estimate ($\widehat{h^2}$) in the null model without variants to determine η.

In the context of GWAS and fine-mapping, we are only interested in variant effects ($\boldsymbol{a}$). Single-marker association is considered as a special case of SNP-set association with set size equal to 1. SNP weighting via $\boldsymbol{A}$ in model (3.1) matters in SNP-set tests and fine-mapping where multiple variants are modeled. A key to improving statistical power of SNP-set tests is properly specifying differential weights for SNPs given their MAFs (Ionita-Laza *et al*, 2013; Wu *et al*, 2011) or functional annotations (Hao *et al*, 2018). Note that a SNP weighting scheme is generic and can be used in any association test methods, such as likelihood ratio test, score test, or use of Bayes factor as in this study. Additionally, weighting variants via $\boldsymbol{A}$ is equivalent to scaling genotypes by square root of corresponding weights (see Appendix A for the proof); for example, using standardized genotypes and setting $\boldsymbol{A} = \boldsymbol{I}$ in model (3.1) is equivalent to using 0/1/2 (additive genotype coding) and setting $A_{ii} = 1/(2 \times \mathrm{MAF}_i \times (1 - \mathrm{MAF}_i))$.

Most existing GWAS approaches, like EMMAX (Kang *et al*, 2010), GEMMA (Zhou and Stephens, 2012), BOLT-LMM (Loh *et al*, 2015b), and SKAT (Wu *et al*, 2011), assume that residuals are independently and identically distributed, that is, ***R***=***I*** as in model (3.1). This generally works well for human phenotypes. However, indirect phenotypes (such as breeding values) are often used in animal and plant GWAS, where modeling their reliability is sometimes critical (see Chapter 4 for our cattle GWAS). We can use $R_{ii} = 1/r^2 - 1$, where r^2 is reliability (VanRaden, 2008). Modeling ***R*** in eigendecomposition methods is straightforward. MMAP has this function (https://mmap.github.io/).

Next, we describe how to efficiently compute *P*(*D*|*M*) (data *D*, and model *M* regarding variant inclusion) by integrating out σ_e^2 based on model (3.1).

Computation of *P*(*D*|*M*)

For any model *M* that defines a variant set to be included in model (3.1), let $\boldsymbol{Z}_M$ (a subset of ***Z***) represent the genotypes of the corresponding variant set. Given $D = \{\mathbf{y}, \mathbf{X}, \mathbf{Z}, \mathbf{G}, \mathbf{R}\}$, we have $P(D \mid M) = P(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, \mathbf{G}, \mathbf{R}, M)P(\mathbf{X}, \mathbf{Z}, \mathbf{G}, \mathbf{R} \mid M)$. Assuming that variant genotypes alone do not contain information about model *M*, $P(\mathbf{X}, \mathbf{Z}, \mathbf{G}, \mathbf{R} \mid M) = P(\mathbf{X}, \mathbf{Z}, \mathbf{G}, \mathbf{R}) = C$ remains constant for any model *M*.

$$
\begin{aligned}
P(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, \mathbf{G}, \mathbf{R}, M) &= P(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}_M, \mathbf{G}, \mathbf{R}) \\
&= \int P(\mathbf{y} \mid \mathbf{V}, \sigma_e^2) P(\sigma_e^2) d\sigma_e^2 \\
&= \int \left(2\pi\sigma_e^2\right)^{-\frac{n}{2}} |\mathbf{V}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}\sigma_e^{-2}\right) \left(\sigma_e^2\right)^{-1} d\sigma_e^2 \\
&= \frac{\Gamma(\alpha)}{\beta^\alpha} (2\pi)^{-\frac{n}{2}} |\mathbf{V}|^{-\frac{1}{2}}
\end{aligned}
$$

where $\mathbf{V} = \varphi \mathbf{X}\mathbf{X}' + \gamma \mathbf{Z}_M \mathbf{A}_M \mathbf{Z}'_M + \eta \mathbf{G} + \mathbf{R}$, $\alpha = n/2$, and $\beta = \mathbf{y}'\mathbf{V}^{-1}\mathbf{y}/2$.

Thus,

$$\log P(D \mid M) = \log \Gamma(\alpha) - \alpha \log \beta - \frac{n}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{V}| + \log C. \qquad (3.2)$$

Evaluating $\log P(D \mid M)$ involves computation of the determinant and inverse of matrix *V*. The computations of $|\mathbf{V}|$ and $\mathbf{V}^{-1}$ can be eased by applications of Sylvester's determinant identity and Woodbury matrix identity, respectively. We have $|\mathbf{V}| = |\mathbf{W}| |\mathbf{I} + \mathbf{K}'\mathbf{W}^{-1}\mathbf{K}|$ and $\mathbf{V}^{-1} = \mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{K}\left(\mathbf{I} + \mathbf{K}'\mathbf{W}^{-1}\mathbf{K}\right)^{-1} \mathbf{K}'\mathbf{W}^{-1}$, where $\mathbf{W} = \eta \mathbf{G} + \mathbf{R}$ and $\mathbf{K} = \left[\sqrt{\varphi}\mathbf{X} \ \sqrt{\gamma}\mathbf{Z}_M \mathbf{A}_M^{1/2}\right]$. With the two equations, we only need to compute the determinant and inverse of *W* once and use them for all variants or variant-sets. Inverse of *V* involves a matrix inversion that has a dimension equal to total number of covariates and included variants (usually much smaller than sample size *n*).

Alternatively, we can use a linear transformation to ease the evaluation of $\log P(D \mid M)$. Let $\mathbf{W} = \mathbf{L}\mathbf{L}'$ (Cholesky decomposition) and $T(\mathbf{v}) = \mathbf{L}^{-1}\mathbf{v}$. With the linear transformation *T*(**v**), model (3.1) becomes

$$\begin{gathered} \mathbf{L}^{-1}\mathbf{y} = \mathbf{L}^{-1}\mathbf{X}\mathbf{b} + \mathbf{L}^{-1}\mathbf{Z}\mathbf{a} + \mathbf{L}^{-1}(\mathbf{g} + \mathbf{e}) \\ \mathbf{y}^* = \mathbf{X}^*\mathbf{b} + \mathbf{Z}^*\mathbf{a} + \mathbf{e}^* \\ \mathbf{b} \sim N(0, \varphi\sigma_e^2\mathbf{I}) \\ \mathbf{a} \sim N(0, \gamma\sigma_e^2\mathbf{A}) \\ \mathbf{e}^* \sim N(0, \sigma_e^2\mathbf{I}) \\ P(\sigma_e^2) \propto 1/\sigma_e^2 \end{gathered} , \qquad (3.3)$$

where $\mathbf{y}^* = \mathbf{L}^{-1}\mathbf{y}$, $\mathbf{X}^* = \mathbf{L}^{-1}\mathbf{X}$, and $\mathbf{Z}^* = \mathbf{L}^{-1}\mathbf{Z}$. Given $D^* = \{\mathbf{y}^*, \mathbf{X}^*, \mathbf{Z}^*\}$, we get

$$P(\mathbf{y}^* \mid \mathbf{X}^*, \mathbf{Z}^*, M) = \frac{\Gamma\left(\alpha^*\right)}{\left(\beta^*\right)^{\alpha^*}} \left(2\pi\right)^{-\frac{n}{2}} \left|\mathbf{V}^*\right|^{-\frac{1}{2}}, \tag{3.4}$$

and

$$P(D^* \mid M) = P(\mathbf{y}^* \mid \mathbf{X}^*, \mathbf{Z}^*, M) P(\mathbf{X}^*, \mathbf{Z}^* \mid M) = P(\mathbf{y}^* \mid \mathbf{X}^*, \mathbf{Z}^*, M) C^*, \tag{3.5}$$

where $\mathbf{V}^* = \varphi \mathbf{X}^* \mathbf{X}^{*\prime} + \gamma \mathbf{Z}_M^* \mathbf{A}_M \mathbf{Z}_M^*{}' + \mathbf{I}$, $\alpha^* = n/2$, $\beta^* = \mathbf{y}^{*\prime} \mathbf{V}^{*-1} \mathbf{y}^* / 2$, and C^* is a constant comparable to C. It is easy to show

$$P(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, \mathbf{G}, \mathbf{R}, M) / P(\mathbf{y}^* \mid \mathbf{X}^*, \mathbf{Z}^*, M) = \left|\mathbf{L}\right|^{-1}.$$

Thus,

$$\log P(D \mid M) = \log P(\mathbf{y}^* \mid \mathbf{X}^*, \mathbf{Z}^*, M) - \log\left|\mathbf{L}\right| + \log C. \tag{3.6}$$

Evaluating $\log P(D^* \mid M)$ requires the determinant and inverse of matrix $\boldsymbol{V}^*$. Let $\mathbf{K}^* = \left[\sqrt{\varphi}\mathbf{X}^* \;\; \sqrt{\gamma}\mathbf{Z}_M^* \mathbf{A}_M^{*1/2} \right]$. Based on Sylvester's determinant identity and Woodbury matrix identity, we get $\left|\mathbf{V}^*\right| = \left|\mathbf{I} + \mathbf{K}^{*\prime}\mathbf{K}^*\right|$ and $\mathbf{V}^{*-1} = \mathbf{I} - \mathbf{K}^* \left(\mathbf{I} + \mathbf{K}^{*\prime}\mathbf{K}^*\right)^{-1} \mathbf{K}^{*\prime}$, respectively.

In proper software implementation, use of equation (3.6) does not necessarily increase speed or reduce memory usage compared to direct use of equation (3.2). However, the linear transformation of model (3.1) to model (3.3) illustrates how we can calculate scaled Bayes factor with model (3.1) and get its null distribution (Zhou and Guan, 2017b).

Null distribution of Bayes factor

Zhou and Guan (2017b) studied the null distribution of Bayes factor (H_0: $\boldsymbol{a}=\mathbf{0}$) in a linear regression model identical to model (3.3). Their theory holds true as long as $\varphi \to \infty$ in model (3.3). We here extend it to model (3.1).

In association tests, we compare a model of interest (M_1) to the null model without variants (M_0). Bayes factor can accordingly be computed for model (3.1) and D by

$$\log \mathrm{BF}_D\left(M_1 : M_0\right) = \log P(D \mid M_1) - \log P(D \mid M_0) ,$$

and for model (3.3) and D^* by

$$\log \mathrm{BF}_{D^*}\left(M_1 : M_0\right) = \log P(D^* \mid M_1) - \log P(D^* \mid M_0) ,$$

respectively. Based on equations (3.5) and (3.6), we can further get

$$\begin{aligned} &\log \mathrm{BF}_D\left(M_1 : M_0\right) \\ &= \log \mathrm{BF}_{D^*}\left(M_1 : M_0\right) \\ &= \log P(\mathbf{y}^* \mid \mathbf{X}^*, \mathbf{Z}^*, M_1) - \log P(\mathbf{y}^* \mid \mathbf{X}^*, \mathbf{Z}^*, M_0) \end{aligned} . \tag{3.7}$$

That is, for any D corresponding to D^*,

$$\mathrm{BF}_D = \mathrm{BF}_{D^*} . \tag{3.8}$$

Because any D is uniquely mapped to D^* and vice versa, BF_D must have the same null distribution as BF_{D^*}. This can be illustrated by using permutation of $\boldsymbol{Z}$ ($\boldsymbol{Z}^*$) (switching individual labels) to create their null distributions. Any permuted D (D_p) is uniquely mapped to a permuted D^* (D_p^*) and vice versa, and $BF_{D_p} = BF_{D_p^*}$. Thus, BF_D has the same null distribution as BF_{D^*}.

Define $\mathbf{P} = \mathbf{I} - \mathbf{X}^{*}\left(\mathbf{X}^{*\prime}\mathbf{X}^{*}\right)^{-1}\mathbf{X}^{*\prime}$, $\mathbf{T} = \mathbf{PZ}^{*}$, and $\mathbf{H} = \mathbf{T}\left(\mathbf{T}'\mathbf{T} + \gamma^{-1}\mathbf{A}^{-1}\right)^{-1}\mathbf{T}'$. Let $\left(\lambda_1, \cdots, \lambda_p\right)$ be eigenvalues of $\boldsymbol{H}$ in descending order. According to Zhou and Guan (2017b),

$$2\log \mathrm{BF} = \sum_{i=1}^{p} \lambda_i Q_i + \sum_{i=1}^{p} \log\left(1 - \lambda_i\right) \text{ with } Q_i \sim \chi_1^2. \tag{3.9}$$

Thus, p-value for $\log \mathrm{BF}$ (H_0: $\boldsymbol{a}=\mathbf{0}$) can be computed by evaluating a weighted sum of chi-squared random variables. We can also compute scaled Bayes factors (Zhou and Guan, 2017b):

$$\begin{aligned} &2\log \mathrm{sBF} \\ &\overset{\text{def}}{=} 2\log \mathrm{BF} - \mathrm{E}_0\left(2\log \mathrm{BF}\right) \\ &= 2\log \mathrm{BF} - \sum_{i=1}^{p} \lambda_i - \sum_{i=1}^{p} \log\left(1 - \lambda_i\right) = \sum_{i=1}^{p} \lambda_i \left(Q_i - 1\right) \end{aligned}, \tag{3.10}$$

where E_0 is the expectation under the null. A large value of γ makes λ_i close to 1, thus resulting in loss of significance in computation of $\log\left(1 - \lambda_i\right)$. To avoid the loss, we instead compute the eigenvalues of $\gamma\mathbf{A}^{1/2}\mathbf{T}'\mathbf{T}\mathbf{A}^{1/2}$ ($\lambda_1^{*}, \cdots, \lambda_p^{*}$ in descending order) and use the relationship between λ_i and λ_i^{*}, $1 - \lambda_i = 1/\left(\lambda_i^{*} + 1\right)$. Furthermore, we use singular value decomposition of $\mathbf{A}^{1/2}\mathbf{T}'$ instead of the eigendecomposition of $\mathbf{A}^{1/2}\mathbf{T}'\mathbf{T}\mathbf{A}^{1/2}$ to improve computation of very small eigenvalues.

When doing association tests or fine-mapping with pedigree data, model (3.1) is used to include a polygenic term to control population structure. The linear transformation $T(\mathbf{v})$ is used to transform phenotypes, covariates, and genotypes. Bayes factor and its null distribution are then computed using the transformed data D^{*} based on model (3.3). The Bayes factor and null distribution are exactly the ones for model (1) with the original data *D*.

Note that $\varphi \rightarrow \infty$ is required. A large value (e.g., 1E8) suffices, when using equations (3.7) and (3.4) to compute Bayes factors. Model M_0 does not include any variants, so the terms for variants need to be removed when computing $P(\mathbf{y}^* \mid \mathbf{X}^*, \mathbf{Z}^*, M_0)$. Alternatively, we can use the following equation to compute Bayes factor (Zhou and Guan, 2017b),

$$\begin{aligned} \log \mathrm{BF} = -\frac{1}{2}\log\left|\gamma \mathbf{A}\right| - \frac{1}{2}\log\left|\mathbf{T}'\mathbf{T} + \gamma^{-1}\mathbf{A}^{-1}\right| \\ + \frac{n}{2}\log\left(\mathbf{y}^{*\prime}\mathbf{P}\mathbf{y}^*\right) - \frac{n}{2}\log\left(\mathbf{y}^{*\prime}\mathbf{P}\mathbf{y}^* - \mathbf{y}^{*\prime}\mathbf{H}\mathbf{y}^*\right) \end{aligned}, \qquad (3.11)$$

where **T**, **P** and **H** are the same as defined previously.

Single-marker/SNP-set association

Based on our derivation above, we can readily compute Bayes factor for a variant set versus the null (equation (3.11)) and corresponding *p*-value (equation (3.9)). For single-marker association tests, the *p*-value associated with Bayes factor (p_B) is asymptotically equal to that from likelihood ratio test (p_F) for the corresponding linear regression, while for SNP-set tests, p_B is generally not equal to p_F as p_B depends on γ. Zhou and Guan (2017b) also showed that p_B is well-calibrated even when the sample size is as small as a few hundred, and the calibration is better than the p_F at very small values.

Evaluating a weighted sum of chi-squared random variables is required to compute p_B for SNP-set association tests, for which we implement saddlepoint approximation (Kuonen, 1999). This method is fast and is accurate in the upper tail, which is sufficient for use in GWAS.

Besides *p*-value, one can use Bayes factor or scaled Bayes factor (sBF) to rank variants besides *p*-value in single-marker association. Scaled Bayes factor is proposed by (Zhou and Guan, 2017). Its definition, computation and null distribution are shown by equation (3.10). Compared to BF, sBF has a few desirable properties (Zhou and Guan, 2017b). For multiple single-marker tests, it has a propensity to assign a larger value to the test that carries more evidence or has a larger power. Simply speaking, among markers with equal *p*-values, the ones with a larger sBF are more appealing. Note that sBF for a SNP set depends on set size and correlations between the SNPs, so sBF is not suitable for ranking SNP-set tests.

Fine-mapping

Fine-mapping is basically model selection problem. We explore the vast model space to find models with highest probability.

Forward selection

We aim to identify independent association signals within a region by forward selection and to assign a posterior probability of causality (PPC) to each variant. Following the first method by Huang *et al* (2017), our fine-mapping approach includes three steps: forward selection (Foster and George, 1994) to add independent signals in the model, repositioning signals, and generating credible variant set for each signal. Though our approach uses the same framework as Huang *et al* (2017), there are a few notable differences (Table 3.1). Specifically, we provide a fast, general-purpose software tool for fine-mapping complex traits, while they only provided R scripts fitting for disease data sets like theirs.

We set $\varphi = \gamma = 1E8$ in model (3.1) for fine-mapping by forward selection, which enables easy calculation of *p*-value for a newly added variant conditioning on variants already added. When existing covariates (including variants that have been added) have an infinite value for φ (φ=1E8 suffices) and design matrix $\boldsymbol{X}^*$, adding variant *i* with transformed genotypes $\boldsymbol{Z}_i^*$ results in:

$$2\log \mathrm{BF} = \lambda_i Q - \log\left(1 + \mathbf{T}_i' \mathbf{T}_i \gamma\right),$$

with $\mathbf{T}_i = \mathbf{P}\mathbf{Z}_i^*$, $\lambda_i = \mathbf{T}_i' \mathbf{T}_i \Big/ \left(\mathbf{T}_i' \mathbf{T}_i + 1/\gamma\right)$, and $Q \sim \chi_1^2$, which is just a special case of equation (3.9). Note that the null model includes variants already added. We set $\gamma = 1E8$ for all variants in fine-mapping, so we get $\lambda_i = 1$ for any variant *i*. Therefore, *p*-value can be easily computed because we have $\left(2\log \mathrm{sBF} + 1\right) \sim \chi_1^2$.

We use Bonferroni threshold (Foster and George, 1994) as stopping criterion in forward selection; that is, forward selection stops when $\left(2\log \mathrm{sBF} + 1\right) < 2\log m_{\mathrm{eff}}$, where m_{eff} is efficient number of independent variants calculated using the method by Li and Ji (2005). Suppose that we select *p* independent signals in forward selection and determine a set of lead variants (S_l) for the *p* signals after repositioning. Then for signal *i* with lead variant (l_{i}), we have a variant set (S_i) containing variants that have substantial LD with l_{i} but weak LD with lead variants in other signals $S_l \setminus \{l_i\}$. Accordingly, we can compute PPC of variant *j* (v_{ij}) in S_i conditioning on $S_l \setminus \{l_i\}$:

$$P(M_i = v_{ij} \mid \mathbf{y}, \mathbf{X}, \mathbf{Z}, S_l \setminus \{l_i\}) = \frac{P(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, M_i = v_{ij}, S_l \setminus \{l_i\}) P(M_i = v_{ij})}{\sum_j P(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, M_i = v_{ij}, S_l \setminus \{l_i\}) P(M_i = v_{ij})}, \qquad (3.12)$$

where $M_i = v_{ij}$ denotes that the causal variant in signal *i* is variant *j* in S_i (i.e. v_{ij}). Efficient computation of $P(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, M)$ is given by equation (3.4). An equal prior for each variant can be used when little prior information is known; that is, $P(M_i = v_{ij}) = 1 \; \forall v_{ij} \in S_i$. We can easily get a credible variant set passing a given confidence level (e.g., 95%) for a signal, by sorting variants in a descending order of PPC and including them in the set from top. We can also calculate PPC of a gene by summing up PPCs of all variants within the gene.

Shotgun stochastic search

There are a total of 2^m models for *m* variants in a region of interest. Enumeration of all models is often infeasible, as *m* is generally more than a few hundred. Let Γ represent the entire model space. We use a shotgun stochastic search (SSS) algorithm to get a subset of Γ, denoted by Γ^*. SSS quickly identifies models with high posterior probability and ignores models with negligible probability (Hans *et al*, 2007), so we anticipate that Γ^*, though much smaller than Γ, contains (almost) all relevant models. Γ^* is accordingly sufficient for follow-up analyses, like prioritization of variants and incorporation of function annotations. As the algorithm has been well described for use in fine-mapping (Benner *et al*, 2016), we skip its details and instead address how to improve its performance.

For any model *M* defining a set of causal variants, its posterior probability is computed by

$$P(M \mid \mathbf{y}, \mathbf{X}, \mathbf{Z}) = \frac{P(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, M) P(M)}{\sum_{M' \in \Gamma^*} P(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, M') P(M')} = \frac{\mathrm{BF}(M : M_0) P(M)}{\sum_{M' \in \Gamma^*} \mathrm{BF}(M' : M_0) P(M')}, \quad (3.13)$$

where M_0 is the null model without variants. (Note $P(\mathbf{X}, \mathbf{Z} \mid M) = P(\mathbf{X}, \mathbf{Z})$ to obtain equation (3.13), assuming that genotypes alone do not contain information about model.) Here we use a standard model prior, $P(M) = (\pi)^{|M|} (1-\pi)^{m-|M|}$ (Hans *et al*, 2007), to induce sparsity, where $|M|$ is model size, and π is a hyperparameter representing the probability that each variant is causal. As this actually induces a binomial prior on model size, the expected model size equals $m\pi$. In fine-mapping, π is often set to $1/m$ so that we expect one causal variant a priori (Benner *et al*, 2016; Chen *et al*, 2015). We further compute posterior inclusion probability (PIP) for variant v_k,

$$\mathrm{PIP}(v_k) = \sum_{M \in \Gamma^*} 1(v_k \in M) P(M \mid \mathbf{y}, \mathbf{X}, \mathbf{Z}). \quad (3.14)$$

The PIP is the marginal posterior probability of causality and measures the relative importance of each variant. PIP has a better performance than ρ-level confidence for prioritizing variants (Chen *et al*, 2015).

A naïve implementation of SSS may fail to discover causal variants even with a very long chain for some LD structures (see examples in Results). The reason may be that a few LD patterns cause the SSS iterations to miss some relevant models. To solve the problem, we introduce simulated annealing to make SSS do more sufficient search. We apply a linear cooling scheme,

$$T_{k+1} = T_k - \Delta T, \quad (3.15)$$

where T_0 and ΔT can be set to 100 and 1 for fine-mapping, respectively, and the final temperature is set to 1. Each temperature is coupled with 10 SSS iterations (though other numbers may apply), except that 100 iterations are used for the final one.

Identical variants (that is, genotype correlation exactly equals 1 or -1) are specially treated in our implementation of SSS. We anticipate that identical variants have the same relevant models; however, SSS cannot guarantee this because it is basically based on random sampling. To solve this problem, after all SSS iterations, we find all models that contain a variant identical to others, and create new models for all possible combinations of identical variants (Fig. 3.1). As a result, identical variants have the same relevant models, which minimizes SSS-induced random errors in computation of PIPs and incorporation of functional annotations. Though being identical with respect to genotypes, their function annotations may be distinct and can be used to distinguish their relevance to phenotypic variation.

Incorporation of functional annotations

We propose an intuitive method to apply differential prior model probabilities to fine-mapping by integrating functional annotation, drawing ideas from a previous study on adjusting significance threshold based on functional annotation in GWAS (Sveinbjornsson *et al*, 2016). This method is readily integrated with our forward selection and SSS approaches and applies to existing software tools (e.g., BIMBAM, CAVIARBF, FINEMAP).

Here we only consider categorical functional annotations. Let $\boldsymbol{c}$ represent functional annotation categories of all variants in a locus. Assuming genotypes are independent of functional annotations, we can obtain:

$$P(\mathbf{y},\mathbf{X},\mathbf{Z},M,\mathbf{c}) = P(\mathbf{y}\mid\mathbf{X},\mathbf{Z},M)P(\mathbf{X},\mathbf{Z}\mid M)P(\mathbf{c}\mid M)P(M)\,. \qquad (3.16)$$

Equation (3.13) correspondingly becomes

$$P(M\mid\mathbf{y},\mathbf{X},\mathbf{Z},\mathbf{c}) = \frac{P(\mathbf{y}\mid\mathbf{X},\mathbf{Z},M)P(M)P(\mathbf{c}\mid M)}{\sum_{M'\in\Gamma^*} P(\mathbf{y}\mid\mathbf{X},\mathbf{Z},M')P(M')P(\mathbf{c}\mid M')}\,. \qquad (3.17)$$

The standard model prior, $P(M) = (\pi)^{|M|}(1-\pi)^{m-|M|}$, is used as in equation (3.13). Incorporation of functional annotation is done in $P(\mathbf{c}|M)$. For a functional annotation with several categories, we define two categorical distributions (one with parameter ***p*** and the other ***q***) and denote the probability of a causal variant being of category c as p_c and the probability of a non-causal variant being of category c as q_c. Assuming that variants are independent of one another with respect to functional annotation, we can compute $P(\mathbf{c}|M)$ for any model M (a set of causal variants) in a locus by taking samples from the two categorical distributions:

$$P(\mathbf{c}\mid M) = \prod_{v\in M} p_{c_v} \prod_{v\notin M} q_{c_v} \propto \prod_{v\in M} p_{c_v}/q_{c_v}\,, \qquad (3.18)$$

where v is a variant in the locus, and c_v denotes its category.

We estimate ***q*** with the genome-wide frequencies of the categories, as in (Sveinbjornsson *et al*, 2016). To estimate ***p***, we can use all the available independent loci in fine-mapping (let Γ_i^* represent SSS model space for locus i):

$$L(\boldsymbol{p}\mid D) = \prod_i P(\mathbf{y},\mathbf{X},\mathbf{Z},\mathbf{c}) \propto \prod_i \sum_{M\in\Gamma_i^*} P(\mathbf{y}\mid\mathbf{X},\mathbf{Z},M)P(M)P(\mathbf{c}\mid M)\,. \qquad (3.19)$$

Taking equation (3.18) into equation (3.19), we obtain the maximum likelihood estimates (MLEs) of ***p*** using the Nelder–Mead method. By taking the estimates of ***p*** and ***q*** to equations (3.18) and (3.17), we get posterior model probabilities with incorporation of function annotation. As shown above, our method is actually an empirical Bayes method.

Note $P(\mathbf{y}|\mathbf{X},\mathbf{Z},M)P(M) \propto P(M|\mathbf{y},\mathbf{X},\mathbf{Z})$ as shown by equation (3.13). Thus, taking equation (3.18), we rewrite equations (3.17) and (3.19) as

$$P(M \mid \mathbf{y},\mathbf{X},\mathbf{Z},\mathbf{c}) = \frac{P(M \mid \mathbf{y},\mathbf{X},\mathbf{Z})\prod_{v\in M} p_{c_v}/q_{c_v}}{\sum_{M'\in\Gamma^*} P(M' \mid \mathbf{y},\mathbf{X},\mathbf{Z})\prod_{v\in M'} p_{c_v}/q_{c_v}} \tag{3.20}$$

and

$$L(\boldsymbol{p} \mid D) \propto \prod_{i}\sum_{M\in\Gamma_i^*} P(M \mid \mathbf{y},\mathbf{X},\mathbf{Z})\prod_{v\in M} p_{c_v}/q_{c_v}\ , \tag{3.21}$$

respectively, where $P(M|\mathbf{y},\mathbf{X},\mathbf{Z})$ is the posterior model probabilities computed without functional annotation. These two new equations suggest an easy-to-use procedure to integrate functional annotation with fine-mapping, which includes three separate steps: i) computing posterior model probabilities $P(M|\mathbf{y},\mathbf{X},\mathbf{Z})$ without functional annotation based on equation (3.13), ii) estimating $\boldsymbol{q}$ with the genome-wide frequencies and taking $P(M|\mathbf{y},\mathbf{X},\mathbf{Z})$ to equation (3.21) to estimate $\boldsymbol{p}$, and iii) taking $P(M|\mathbf{y},\mathbf{X},\mathbf{Z})$ and the estimates of $\boldsymbol{p}$ and $\boldsymbol{q}$ to equation (3.20) to obtain posterior model probabilities $P(M|\mathbf{y},\mathbf{X},\mathbf{Z},\mathbf{c})$ with incorporation of functional annotation and subsequently computing PIPs. Note that we only need to compute the first step once, even when we try many different functional annotations in fine-mapping. This feature makes our approach easier to use compared with PAINTOR (Kichaev *et al*, 2014) and CAVIARBF (Chen *et al*, 2016).

In the derivation above, we address the use of SSS outputs. In fact, this method is also applicable to our forward selection approach. Forward selection identifies independent association signals, and it is usually safe to assume that there is only one causal variant in each signal. Therefore, forward selection outputs can be considered as models of size 1, just a special case of SSS outputs. However, we must be aware that for a locus with multiple signals, the forward selection approach outputs posterior probabilities of variants in a signal conditional on lead variants of other signals. To integrate functional annotation with forward selection outputs, we use the following approximation:

$$P(M_i = v_{ij} \mid \mathbf{y}, \mathbf{X}, \mathbf{Z}) \doteq P(M_i = v_{ij} \mid \mathbf{y}, \mathbf{X}, \mathbf{Z}, S_l \setminus \{l_i\}), \tag{3.22}$$

in which all the denotations are the same as equation (3.12). To make the approximation effective, we may remove a signal when incorporating function annotation into forward selection outputs, if the variants in its credible set have high correlation with those in another signal.

Computing the MLEs of $\boldsymbol{p}$ is time-consuming. To gain speedup, we can disregard bottom models, which has little impact on the estimation. In practice, we use only the top models whose cumulative posterior probability passes a threshold (e.g., 0.9).

Besides the use in fine-mapping, our method is also useful for functional enrichment analysis. The enrichment for category c is defined as $E_c = p_c/q_c$ following (Sveinbjornsson *et al*, 2016), for which a value larger than one indicates that causal variants are more enriched in category c than across whole genome. Its estimate is $\widehat{E_c} = \widehat{p_c}/\widehat{q_c}$, and the confidence interval of the estimate is derived by percentile bootstrap.

Time complexity

Table 3.2 lists time complexity of the computations in our method. Basically, BFMAP has a time complexity similar to EMMAX and SKAT for single-marker GWAS and

SNP-set association, respectively. Model size varies for computation of Bayes factors in fine-mapping. The time complexity of computing one model is approximately $O(p^2n + p^3)$ where *p* represents the number of causal variants and is generally small (considering the causal effects that are detectable with sufficient statistical evidence). About *pm* and *tpm* models are computed in forward selection and SSS, respectively, where *t* is the effective number of SSS iterations (generally much smaller than the actual number specified). Thus, our fine-mapping approaches have a time complexity of $O(p^3mn)$ or $O(tp^3mn)$ when polygenic term is not needed to control population structure.

Software

We develop BFMAP with the Eigen 3 C++ library, implementing our methods for single-marker/SNP-set association and fine-mapping. Incorporation of functional annotation into fine-mapping is implemented separately, with the optim() function in R (Team, 2013).

Benchmarking and application

Data sets

We used two real data sets in our analysis. The first one is a dairy cattle data set, which consists of high-density (HD) genotypes of ~300K SNP markers for ~27,000 Holstein bulls. These bulls represent a complex population and have highly reliable breeding values (PTAs) for 35 production, reproduction, and body conformation traits, with average reliability of 0.71 across traits. Imputed sequence genotypes of ~3 million variants are also available for these bulls. This data set has been used and well described in a previous study (VanRaden *et al*, 2017). SNPs with MAF<0.01 were excluded in further analysis.

The second data set is from the Atherosclerosis Risk in Communities (ARIC) study (Investigators, 1989) and consists of imputed genotypes of ~2.5 million SNPs and phenotypes of four lipid profile traits for 9,713 unrelated European Americans. We removed the SNPs (MAF<0.01 or HWE test *p*<1E-6) before further analysis.

In addition, we simulated 308 data sets using the ARIC genotypes on chromosome 22. There are 30,884 SNPs covering 14.4-49.6 Mb on Chr22. We divided the first 30,800 SNPs on Chr22 into 308 continuous groups, each having 100 SNPs. Within each group, we randomly sampled two variants as causal. Effect size was properly assigned to each causal variant so that we had a power range of (0.527, 0.992) to identify the association when using marginal test statistics with significance level 5E-8 (Chen *et al*, 2015). For each of the 308 SNP groups, we summed the effects of two causal SNPs and a random error (sampled from $N(0,1)$) to obtain phenotypes. Consequently, we got 308 data sets, each consisting of genotypes of 100 SNPs and simulated phenotypes for 9713 individuals. The resulting 308 data sets, which represented a variety of LD structures, were used for validating the fine-mapping performance of our approaches.

Single-marker association

We compared BFMAP with several popular software tools for single-marker GWAS, including GEMMA (Zhou and Stephens, 2012), EMMAX (Kang *et al*, 2010), BOLT-LMM (Loh *et al*, 2015b), and MMAP. The dairy cattle HD data were used for this comparison. Only milk was analyzed by all the software tools, while other traits were analyzed by only BFMAP and MMAP. The heritability estimates needed by BFMAP were obtained by MMAP. We set $\gamma = 1E8$ in BFMAP.

We computed the same type of GRMs in BFMAP, EMMAX, GEMMA and MMAP (so called Balding-Nichols matrix in EMMAX). Note that GEMMA uses sample variance of genotypes when building GRM, while BFMAP, EMMAX and MMAP use expected value ($2 \times \text{MAF} \times (1 - \text{MAF})$). In addition, BOLT-LMM uses leave-one-chromosome-out (LOCO) approach. This may result in unexpected problems in some cases.

SNP-set association

We compared BFMAP with SKAT (Wu *et al*, 2011) for SNP-set association. Here we focus on the comparison between the score test in SKAT and the use of Bayes factor in BFMAP, so we use only a simple SNP weighting scheme ($1/(2 \times \text{MAF} \times (1 - \text{MAF}))$) for additive genotype coding (0, 1, or 2). Both the dairy cattle HD data and the human lipid profile data were used for the comparison. We divided cattle and human genomes into non-overlapping 1-Mb and 100-kb segments, respectively, and SNPs in each segment form a SNP set. Accordingly, we obtained 2,521 and 26,543 SNP sets on autosomes for the cattle and human data, respectively. The maximum set size is 245 for the cattle data and 481 for the human data.

While the human population consists of unrelated individuals, the dairy cattle population has a complex population structure. SKAT needs the null model for the latter case (basically, a linear mixed model) which is computed by EMMA (Kang *et al*, 2008). In contrast, BFMAP needs the heritability whose estimate is computed by MMAP.

Fine-mapping

We simulated 308 data sets (described above) and used them to demonstrate the fine-mapping performance of BFMAP. Only CAVIARBF (Chen *et al*, 2015) was used as a benchmark, as this software can be considered ideal for our small data sets. We ran CAVIARBF with options *-t 0 -a 0.1 -c 2* and *-p 0* for computing Bayes factors and model search, respectively. Since use of summary statistics in CAVIARBF is equivalent to use of standardized additive genotypes in model (3.1), we used $\gamma = 0.01$ and $A_{ii} = 1/(2 \times \mathrm{MAF}_i \times (1 - \mathrm{MAF}_i))$ for additive genotype input (0, 1, or 2) in BFMAP to make the two software tools compute equivalent models. Additionally, we set the maximum number of causal variants to five for SSS in BFMAP, as this setting is generally reasonable in real data analysis. However, setting *-c 5* often results in infeasible computation in CAVIARBF.

Besides the benchmarking, we applied BFMAP to the imputed sequence data of 27K Holstein bulls to demonstrate incorporation of functional annotation. We fine-mapped 13 loci associated with milk yield and incorporated SnpEff-inferred variant impacts (Cingolani *et al*, 2012) into the fine-mapping. There were two few high-impact variants in the 13 loci, so we merged them with moderate-impact ones. Consequently, the functional annotation has three categories, i.e., moderate, low, and modifier.

Results

Single-marker association

We compare BFMAP with MMAP, GEMMA, EMMAX, and BOLT-LMM in terms of *p*-value and computational efficiency.

***P*-value**

As shown in Fig. 3.2A, BFMAP generates the same *p*-values as EMMAX for milk GWAS, while both the software tools have slightly larger *p*-values than MMAP and GEMMA especially at the tail. A tiny difference was also observed between MMAP and GEMMA, which may result from the aforementioned difference in GRM. We further analyzed other 34 dairy cattle traits using BFMAP and MMAP. BFMAP has the same genomic control factor as MMAP for each of the 35 dairy cattle traits (Fig. 3.2B). The two tools generate largely the same *p*-values in GWAS for all the 35 traits except that BFMAP has a slight deflation at the tail for milk, fat, fat percentage, and protein percentage (Fig. 3.3).

We further analyzed milk, fat, fat percentage and protein percentage with the LOCO approach for Chr14 using BFMAP and MMAP. Though BFMAP still has a little deflation at the tail compared to MMAP, the deflation is slightly reduced by LOCO compared to the use of GRM built with all markers (see Fig. 3.4 for the analysis on milk). Additionally, BOLT-LMM, which automatically uses LOCO, gives overall similar results to BFMAP, but has considerable deflation at the tail. However, LOCO is actually not applicable to the dairy cattle data. As shown by Fig. 3.4C, LOCO results in severe inflation across the whole chromosome, making the result hardly interpretable.

Computational efficiency

Table 3.3 lists the time cost of the five software tools for the analysis of milk involving 27,158 animals and ~286,000 markers. Among the four GRM-based tools, BFMAP was the fastest and took only 44.7 minutes, which was 3.2, 7.4 and 16.5 times as fast as MMAP, GEMMA and EMMAX, respectively. It is not straightforward to further compare the four tools with BOLT-LMM, in that BOLT-LMM uses LOCO and computes both infinitesimal model association and mixture model association. A simple observation is that it took 9.9 times as long as BFMAP to complete the GWAS.

SNP-set association

We compare BFMAP with SKAT in terms of *p*-value and computational efficiency using the dairy cattle data and the human lipid profile data.

***P*-value**

As shown in Fig. 3.5, both SKAT and BFMAP obtained reasonable results for all the four lipid profile traits in the human data. Setting γ=1E-6 imposed a tiny-effect prior on variants in BFMAP, leading to largely the same results as SKAT for all the traits. When a moderate-effect prior was imposed on variants by setting γ=0.01 in BFMAP, positive association signals were generally inflated compared to those in SKAT (Fig. 3.5).

As for five milk production traits in the dairy cattle data, we had similar observations. BFMAP with a tiny-effect prior (γ=1E-5) resulted in largely the same *p*-values as SKAT for all SNP-set association tests for all the five traits (Figs. 3.6 and 3.7). In contrast, BFMAP with a moderate-effect prior (γ=0.01) inflated positive association signals as compared to SKAT. Overall, the result implies that BFMAP can achieve a higher power than SKAT with proper specification of hyper-parameter γ while being similar to SKAT for controlling false positives.

Computational efficiency

BFMAP is 6-9 times as fast as SKAT for analyzing the human lipid profile traits (Table 3.4) and 3.0 times for analyzing the dairy cattle data (Table 3.5). Note that for the dairy cattle data, the null model is a mixed model. SKAT uses EMMA to compute the null, while BFMAP uses MMAP. Speedup in BFMAP partly results from MMAP which is 2.6 times as fast as EMMA for computing the null.

Fine-mapping

We first compare BFMAP with CAVIARBF using simulated data, and then demonstrate the incorporation of functional annotation in fine-mapping by applying BFMAP to the dairy cattle imputed sequence data.

Fine-mapping accuracy

We compared PIPs (or PPCs for forward selection) of all causal variants computed by BFMAP to those by CAVIARBF. As shown in Figs. 3.8A and 3.8C, SSS had slightly better fine-mapping accuracy than forward selection when we used 1100 SSS iterations without simulated annealing, but both SSS and forward selection missed some causal variants. As we increased the number of SSS iterations to 50K, BFMAP performed much better (Fig. 3.8B), but there were still four causal variants (circled in Fig. 3.8B) with much lower PIPs than expected. We further analyzed the two data sets involving the four causal variants with a longer SSS chain. Even with 500K iterations, SSS still failed to compute correct PIPs for two of the four. In contrast, when simulated annealing was coupled with SSS, BFMAP SSS obtained largely the same PIPs of causal variants as CAVIARBF (Fig. 3.8D).

Incorporation of functional annotation

We fine-mapped milk with BFMAP SSS, and then used the resulting posterior model probabilities to estimate p_c for the three categories of the SnpEff-inferred effect impact (moderate, low, and modifier). We tried three different cumulative posterior probability thresholds for keeping top models (0.8, 0.9, and 0.99) when estimating $\boldsymbol{p}$, and obtained similar estimates (Table 3.6). This suggests that disregarding bottom models has little effect on the estimation of $\boldsymbol{p}$.

We used the estimate of $\boldsymbol{p}$ computed with the threshold 0.9 for the following analysis. About 11.7x and 5.4x enrichment of causal variants were observed in moderate-impact variants and low-impact ones, respectively (Fig. 3.9A). Accordingly, incorporation of this functional annotation into fine-mapping increased PIPs of moderate- and low-impact variants to some extent while reducing PIPs of modifier variants (Fig. 3.9B).

Computational efficiency

Table 3.7 lists the time cost for fine-mapping milk by BFMAP. The SSS with simulated annealing evaluated ~1.8 million distinct models for analyzing 2297 variants in one locus, taking 32.7 minutes with 8 cores on Intel Xeon CPU E5-2680 v2. The forward selection was much faster for analyzing the same locus. Estimating $\boldsymbol{p}$ for the functional annotation involved all the 13 loci, taking 13.1 minutes with one core on Intel Core i7-4790. Re-computing PIPs (based on functional annotation) for all variants in the 13 loci took ~11 minutes. Overall, BFMAP has a reasonable computing speed for fine-mapping.

Discussion

In summary, we propose a unified Bayesian model for single-marker/SNP-set association and fine-mapping and develop an efficient software tool, BFMAP, to deal with both population and pedigree data. Extensive data analyses show that BFMAP achieves a power similar to or higher than existing software tools but is at least a few times faster with respect to single-marker/SNP-set association tests. We also demonstrate that BFMAP performs well for fine-mapping and easily incorporates functional annotation.

In single-marker association tests, we compare BFMAP with MMAP, EMMAX, GEMMA, and BOLT-LMM. BFMAP and EMMAX generate the same *p*-values, because 1) both use the heritability estimate from null model for all SNPs, and 2) the *p*-value associated with Bayes factor for our Bayesian model is asymptotically equal to that from likelihood ratio test (or Wald test) for the corresponding linear regression. MMAP and GEMMA are similar to each other in that both computes exact test statistics, which may gain power for traits influenced by large-effect QTLs compared to EMMAX and BFMAP (Zhou and Stephens, 2012). In theory, the approximation used in BFMAP generally compromises tests for only large-effects QTLs. This is validated by the analyses of 35 dairy cattle traits with MMAP and BFMAP. Compared to MMAP, BFMAP results in a small deflation of *p*-values at the tail for milk, fat, fat percentage and protein percentage (Fig. 3.3). These four traits are well-known for their large-effect causal genes, DGAT1 (Grisart *et al*, 2004) and ABCG2 (Cohen-Zinder *et al*, 2005). For all the other 31 traits, BFMAP has largely the same results as MMAP (Fig. 3.3).

The LOCO approach can well improve the power of GWAS (Listgarten *et al*, 2012; Loh *et al*, 2015b; Yang *et al*, 2014) and partly reduce the deflation at the tail caused by the approximation used in BFMAP. Despite the benefits, LOCO may result in unexpected severe inflation, which is demonstrated by our LOCO analysis on Chr14 for four dairy traits (see Fig 3.4 for milk as an example). There are significant SNPs (P<5E-8) everywhere on the chromosome. We suppose that the use of breeding values instead of direct phenotypes may account for the striking difference from human studies. In addition, BOLT-LMM produces considerably deflated *p*-values at the tail compared to BFMAP (Fig. 3.4A), which may be due to use of the fast approximation similar to GRAMMA-Gamma (Svishcheva *et al*, 2012).

Besides the four software tools compared to BFMAP, there are other software tools based on similar computational and statistical approaches, such as FaST-LMM (Lippert *et al*, 2011) and GCTA (Yang *et al*, 2014). These tools have been well compared with EMMAX, GEMMA or BOLT-LMM in terms of running speed. Among the tools computing exact statistics, MMAP is recommended, because it is clearly the winner with respect to speed. If exact statistic is not required (which is often true in practice), BFMAP is a better choice. When LOCO is preferred, BOLT-LMM is the best choice. Additionally, BFMAP and MMAP can model reliability or accuracy of phenotypic records, which is beneficial or even necessary for analyzing breeding values arising from plant and animal breeding.

In SNP-set tests, BFMAP attains a smaller *p*-value than SKAT for positive loci, when a moderate value (e.g. 0.01) is set to the hyper-parameter (γ). This larger power may be because the prior by setting γ =0.01 is more consistent with the true effect size of the loci. In contrast, setting γ =1E-6 is similar to assuming an infinitesimal genetic architecture. In practice, we can run BFMAP two times (one with small γ value, and the other with moderate γ value) to better model the genetic architecture for a trait of interest and to maximize the power.

We have implemented saddlepoint approximation to evaluate a weighted sum of chi-squared random variables for SNP-set association tests, which produces accurate *p*-values. If a more accurate *p*-value is needed, one can use BFMAP outputs (basically, weights and Bayes factor) to re-compute it by external software tools, e.g., BACH (Zhou and Guan, 2017b).

We demonstrate that some LD structures may hinder sufficient model search of SSS, but that can be overcome by introducing simulated annealing. We notice that FINEMAP produced accurate PIPs with only 100 SSS iterations in a previous study (Benner *et al*, 2016). The difference from our result may result from different simulation procedures: 1) our simulation covers most of chromosome 22, 2) two causal variants in each data set are always within ~100 kb region, and 3) LD pruning is not used in our data. Additionally, we have used standard SNP weighting in BFMAP to make our model equivalent to CAVIARBF. In practice, we can use a different SNP weighting scheme; e.g., using non-standardized SNP genotypes (which assumes that high-MAF SNPs have larger per-SNP heritability than low-MAF ones) may better account for the genetic architecture of some complex traits (like the milk production traits in dairy cattle).

We develop an empirical Bayes method to incorporate functional annotation into fine-mapping. It is actually not only a method for fine-mapping, but also an approach for functional enrichment analysis based on GWAS signals (Fig. 3.9A). Initial BFMAP fine-mapping outputs can be used repeatedly to analyze enrichment patterns of causal variants for many functional annotations. Such enrichment, by definition, is different from enrichment of heritability computed by stratified LD score regression (Finucane *et al*, 2015). The latter one is based on all available markers, while the former one is based on a limited number of QTLs. In addition, our current implementation fits one functional annotation at a time. Enhancing it to model multiple annotations simultaneously seems interesting for further study.

References

Bartlett MS (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44**(3-4)**:** 533-534.

Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**(10)**:** 1493-1501.

Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR *et al* (2015a). An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**(11)**:** 1236-1241.

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C *et al* (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**(3)**:** 291-295.

Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA *et al* (2015). Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**(3)**:** 719-736.

Chen W, McDonnell SK, Thibodeau SN, Tillmans LS, Schaid DJ (2016). Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics. *Genetics* **204**(3)**:** 933-958.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L *et al* (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**(2)**:** 80-92.

Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee JH *et al* (2005). Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res* **15**(7)**:** 936-944.

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR *et al* (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**(11)**:** 1228-1235.

Foster DP, George EI (1994). The Risk Inflation Criterion for Multiple-Regression. *Ann Stat* **22**(4)**:** 1947-1975.

Grisart B, Farnir F, Karim L, Cambisano N, Kim JJ, Kvasz A *et al* (2004). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc Natl Acad Sci U S A* **101**(8)**:** 2398-2403.

Hans C, Dobra A, West M (2007). Shotgun Stochastic search for "Large p" regression. *Journal of the American Statistical Association* **102**(478)**:** 507-516.

Hao X, Zeng P, Zhang S, Zhou X (2018). Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS Genet* **14**(1)**:** e1007186.

Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* **198**(2)**:** 497-508.

Huang H, Fang M, Jostins L, Umicevic Mirkov M, Boucher G, Anderson CA *et al* (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**(7662)**:** 173-178.

Investigators A (1989). The atherosclerosis risk in communit (aric) study: Design and objectwes. *American journal of epidemiology* **129**(4)**:** 687-702.

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X (2013). Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet* **92**(6)**:** 841-853.

Jiang J, Shen B, O'Connell JR, VanRaden PM, Cole JB, Ma L (2017). Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. *BMC Genomics* **18**(1)**:** 425.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB *et al* (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**(4)**:** 348-354.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ *et al* (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**(3)**:** 1709-1723.

Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL *et al* (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**(10)**:** e1004722.

Kuonen D (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**(4)**:** 929-935.

Li J, Ji L (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**(3)**:** 221-227.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011). FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**(10)**:** 833-835.

Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D (2012). Improved linear mixed models for genome-wide association studies. *Nat Methods* **9**(6)**:** 525-526.

Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM *et al* (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**(3)**:** 284-290.

Servin B, Stephens M (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**(7)**:** e114.

Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A, Masson G *et al* (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet* **48**(3)**:** 314-317.

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012). Rapid variance components-based method for whole-genome association analysis. *Nat Genet* **44**(10)**:** 1166-1170.

Team RC (2013). R: A language and environment for statistical computing.

VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**(11)**:** 4414-4423.

VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet Sel Evol* **49**(1)**:** 32.

Wakefield J (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genet Epidemiol* **33**(1)**:** 79-86.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**(1)**:** 82-93.

Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH *et al* (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47**(10)**:** 1114-1120.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR *et al* (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**(7)**:** 565-569.

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM *et al* (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* **43**(6)**:** 519-525.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**(2)**:** 100-106.

Zhou Q, Guan Y (2017). On the Null Distribution of Bayes Factors in Linear Regression. *Journal of the American Statistical Association***:** 1-10.

Zhou X, Stephens M (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**(7)**:** 821-824.

Tables

Table 3.1. Differences between BFMAP and the fine-mapping approach by Huang et al.

	BFMAP	**Huang et al. (Nature 2017)**
Trait type	Single quantitative trait	Multiple disease traits
Calculation of log(*D*|*M*)	Exact	BIC approximation
Calculation of effective number of independent variants	Li & Ji (Heredity 2005)	Recursive method based on correlations between variants
Stopping criterion in forward selection	Bonferroni threshold	Minimum BIC
***P*-value calculation**	Null distribution of log(sBF)	Likelihood ratio test
Software implementation	Fast, general-purpose C++ program	R scripts fitting for their own data
Incorporation with functional annotation	Implemented in R scripts	N.A.

Table 3.2. Time complexity of BFMAP algorithms

Input data	Data transformation	Bayes factor[1]	Null distribution of Bayes factor[1]	GWAS with equal SNP-set size	Forward Selection	SSS
polygenic term is included	$O(n^3 + cn^2 + mn^2)$	$O(s^2n + s^3)$	$O(s^2n)$	$O(n^3 + cn^2 + mn^2 + mns + ms^2)$	$O(n^3 + cn^2 + mn^2 + p^3mn)$	$O(n^3 + cn^2 + mn^2 + tp^3mn)$
polygenic term is not included	$O(cn + mn)$			$O(cn + mns + ms^2)$	$O(cn + p^3mn)$	$O(cn + tp^3mn)$

n: sample size. *c*: number of covariates. *m*: number of markers. *s*: SNP-set size. *p*: number of putative causal variants.

[1] This is the time for one SNP-set with transformed data. Time for computing covariate-related terms is not included, because the computation is done only once and used for all variants.

Table 3.3. Time costs of five software tools for the analysis of milk involving 27,158 animals and ~286,000 markers

Software[1]	**Version**	**GRM (minutes)**[2]	**GWAS (minutes)**[3]
BFMAP		48.9	44.7
MMAP	2017_08_18 (binary)	57.0	142.0
GEMMA	0.97 Guix generic	128.7	330.8
EMMAX	20120210 (Intel binary)	62.8	735.8
BOLT-LMM	v2.3.2 (binary)	N.A.	442.2

[1] All tools were tested using 8 cores of Intel Xeon E5-2680 v2 in the Deepthought2 HPC cluster at the University of Maryland.

[2] MMAP was run with --group_size 10000. EMMAX was run with -M 20.0.

[3] The running time of BFMAP includes 16.2 minutes taken by MMAP to fit the null model. The time of BOLT-LMM includes computation of both infinitesimal-model association statistics and mixture-model association statistics.

Table 3.4. R Time cost for analyzing human lipid profile traits with BFMAP and SKAT

Trait	**N_samples**	**N_markers**	**N_sets**	**Time (minutes)**	
				BFMAP	SKAT[1]
TC	9,156	2,415,449	26,543	31.63	271.94
LDL	9,071	2,415,449	26,543	33.7	205.99
HDL	9,131	2,415,449	26,543	32.18	206.37
TG	9,156	2,415,449	26,543	33.63	204.78

TC: total cholesterol. LDL: low-density lipoprotein. HDL: high-density lipoprotein. TG: triglycerides.

[1] Time cost for TC includes that for getting SNP set data with Generate_SSD_SetID(). The SNP set data were reused for LDL, HDL and TG.

Table 3.5. Time cost for analyzing dairy milk with BFMAP and SKAT

Software[1]	**Computing null model[2]**	**Scanning SNP sets**	**Total time**
BFMAP	88.3	190.6	278.9
SKAT	228.3	608.1	836.4

[1] All tools were tested using 1 core of Intel Xeon E5-2680 v2.

[2] MMAP is used to compute null model for BFMAP. SKAT invokes EMMA to compute null model.

Table 3.6. Number of used models and probability estimates of causal variants being of each category given different cumulative probability thresholds

Cumulative probability threshold	**# loci**	**# models**	***P_c***		
			LOW	MODERATE	MODIFIER
0.8	13	15721	0.111	0.158	0.731
0.9	13	69823	0.114	0.163	0.723
0.99	13	494280	0.119	0.161	0.720

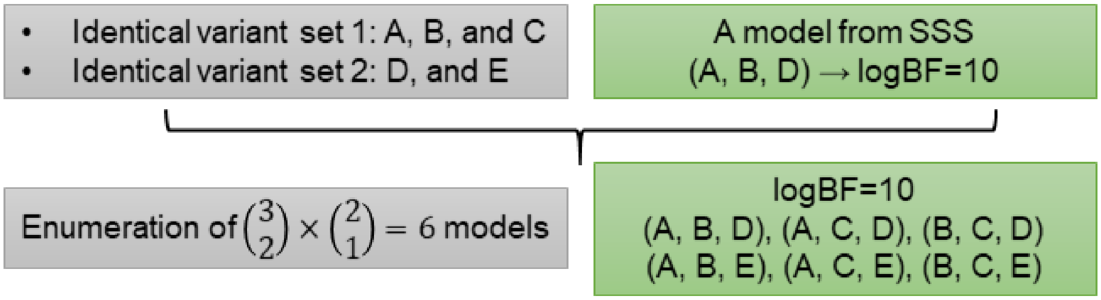
Table 3.7. Running time of BFMAP for fine-mapping dairy milk

Computation	**# animals**	**# loci**	**# variants**	**# models**	**Time cost (minutes)**
SSS[1]	27158	1	2297	1832046	32.7
Forward selection[1]	27158	1	2297	N.A.	1.5
Estimating *p*[2]	N.A.	13	N.A.	69823	13.1
Computing PIP with *p* and *q*[2]	N.A.	13	28728	1335656	11

[1] One locus was analyzed.

[2] All 13 loci were analyzed for incorporating functional annotation into fine-mapping.

Figures



- Identical variant set 1: A, B, and C
- Identical variant set 2: D, and E

A model from SSS
(A, B, D) → logBF=10

Enumeration of $\binom{3}{2} \times \binom{2}{1} = 6$ models

logBF=10
(A, B, D), (A, C, D), (B, C, D)
(A, B, E), (A, C, E), (B, C, E)

Figure 3.1. Processing of models containing identical variants in BFMAP

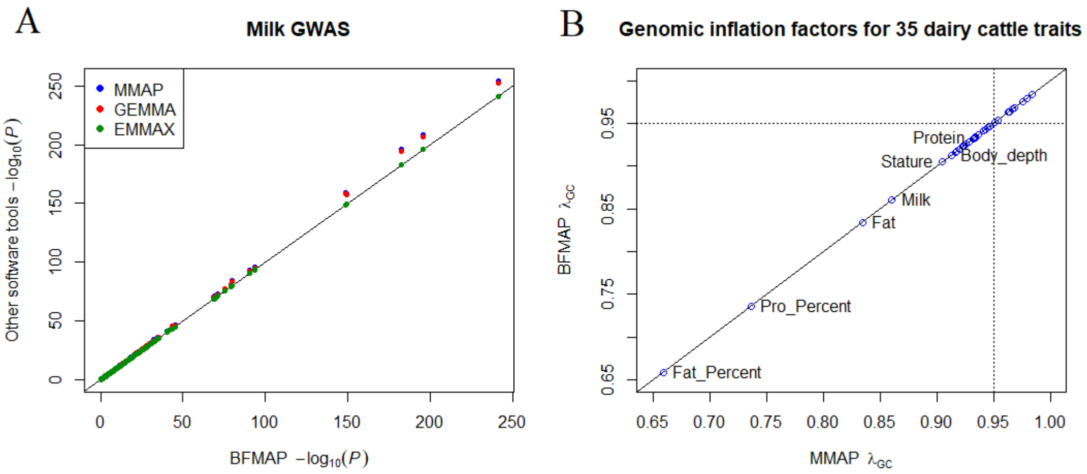


Figure 3.2. Comparison between BFMAP and other software tools in terms of *p*-values and genomic inflation factors. The solid black lines are *y*=*x*.

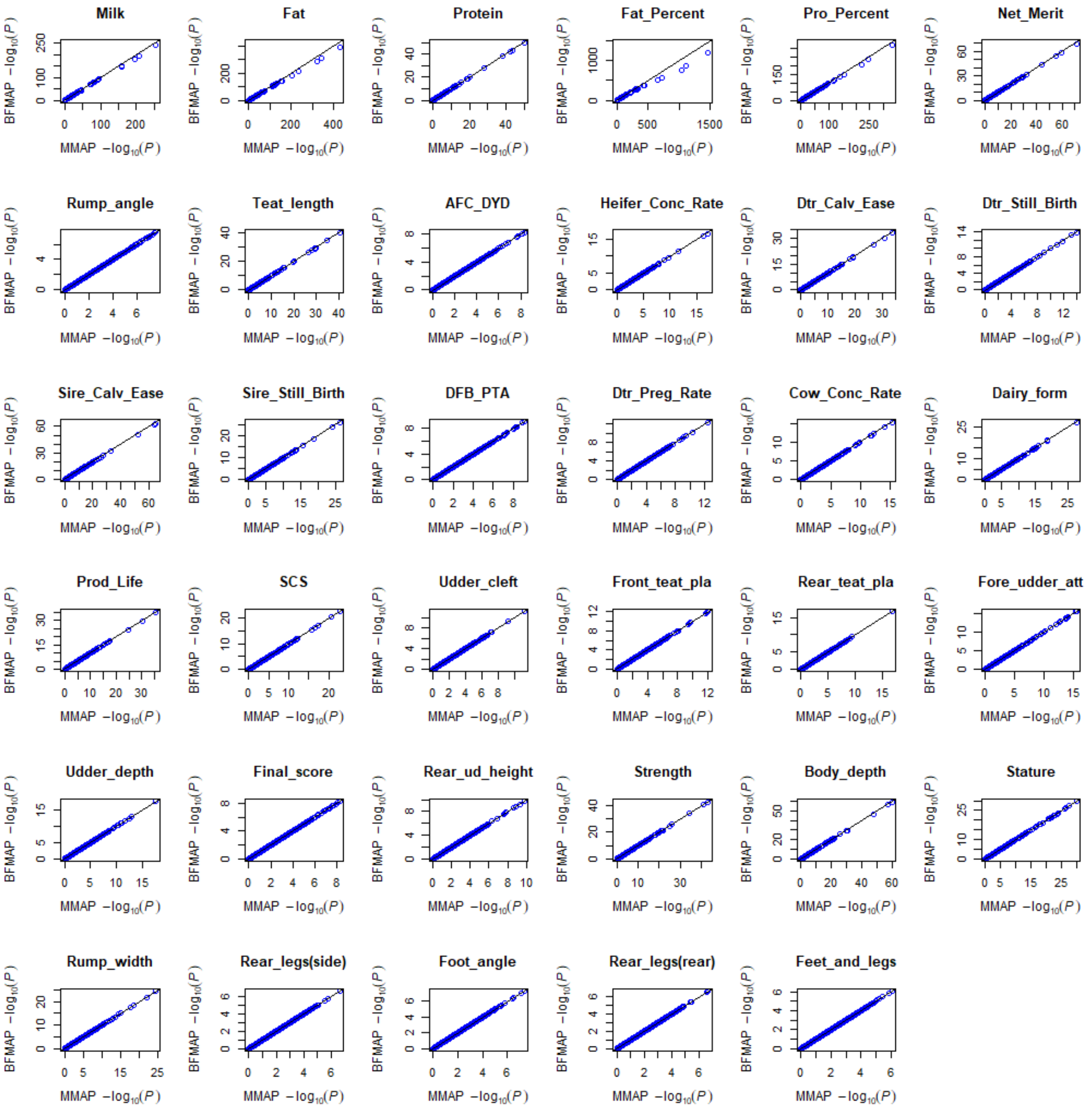


Figure 3.3. Comparison between BFMAP and MMAP in terms of *p*-values for 35 dairy cattle traits.

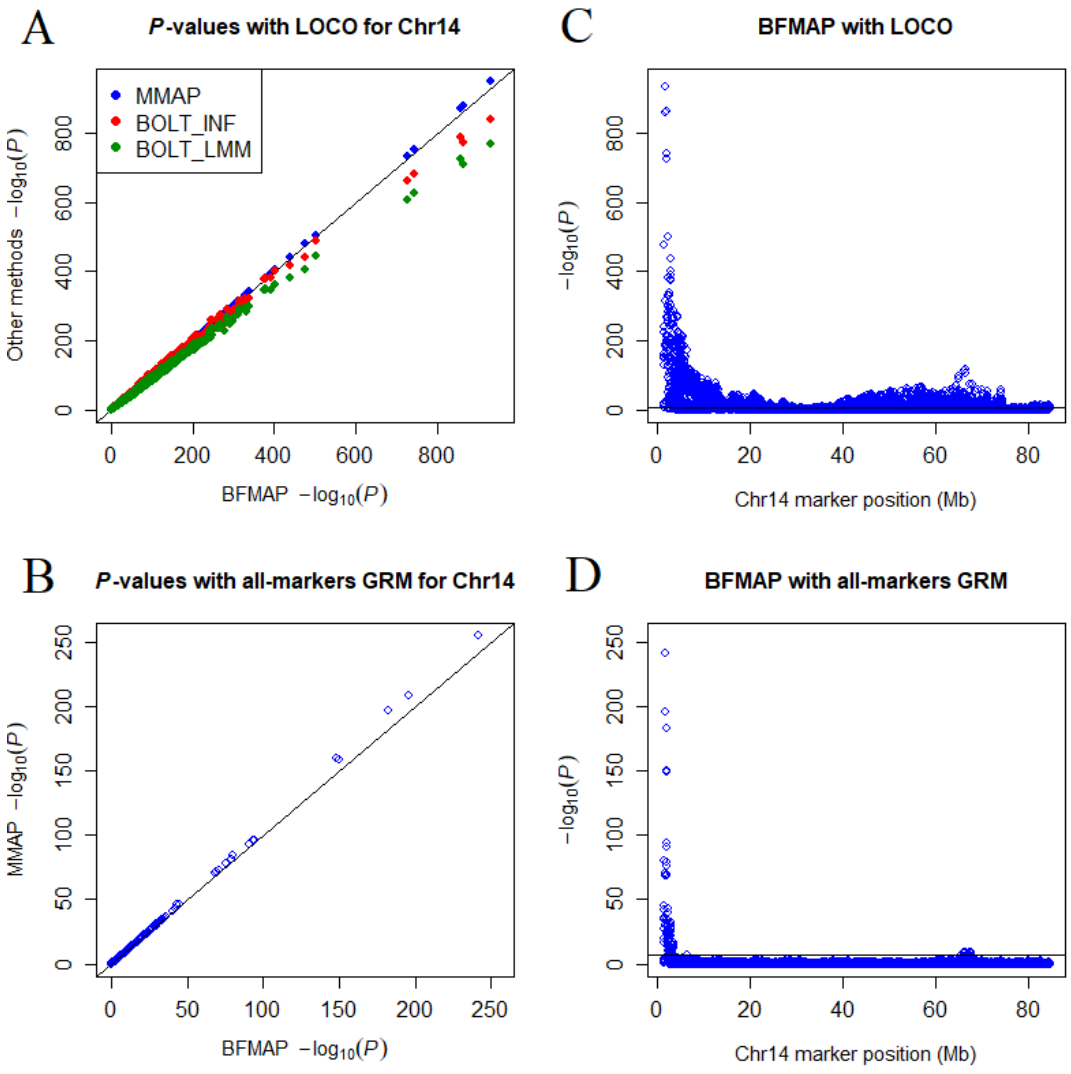


Figure 3.4. The LOCO analysis results of BFMAP, MMAP, and BOLT-LMM for dairy milk. Chromosome 14 is shown here. The black lines in panels A and B are *y*=*x*. The black lines in C and D represent the *p*-value threshold, 5E-8.

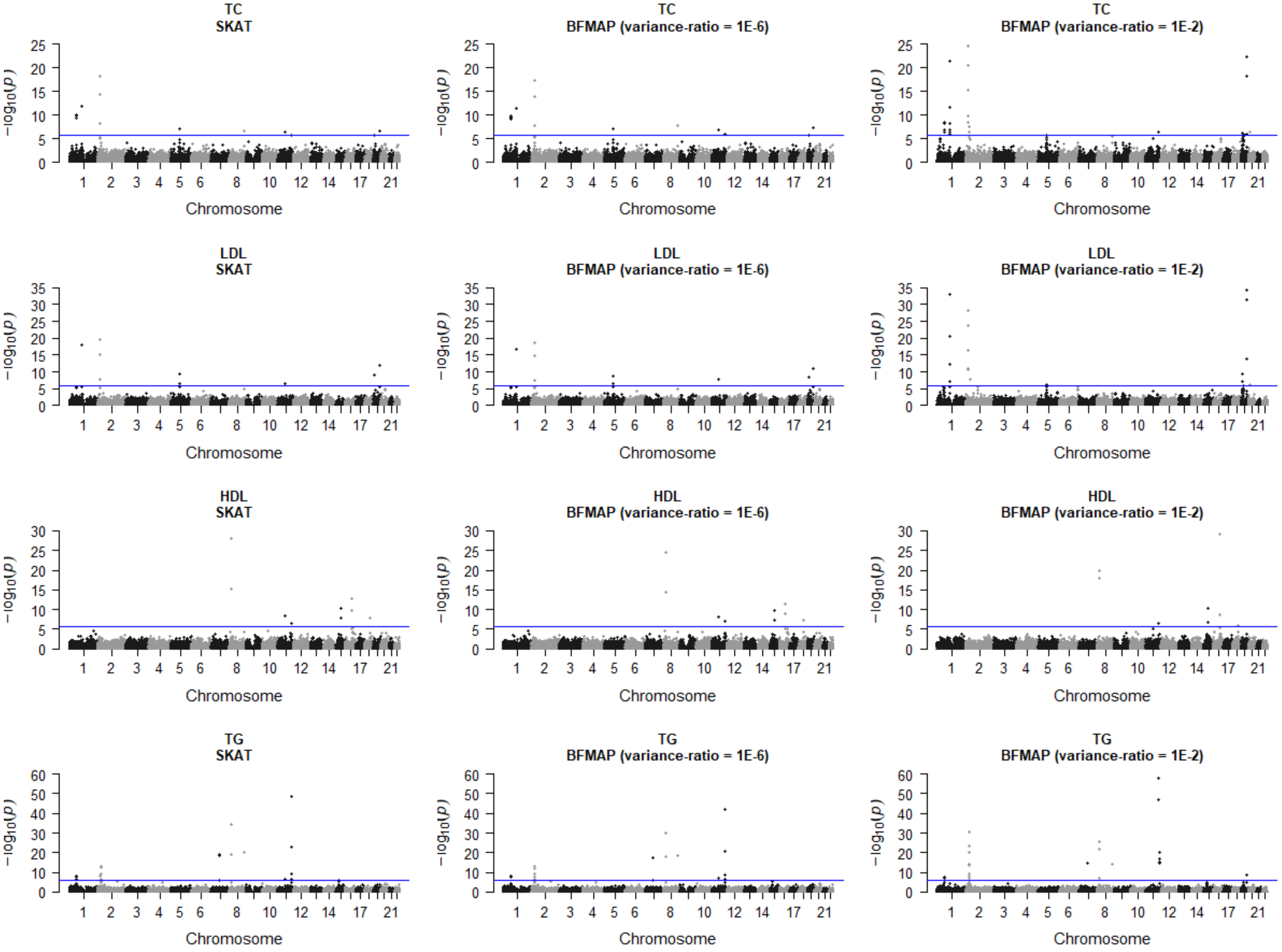


Figure 3.5. SNP-set tests for four human lipid profile traits with BFMAP and SKAT. Variance-ratio means γ in model (3.1).

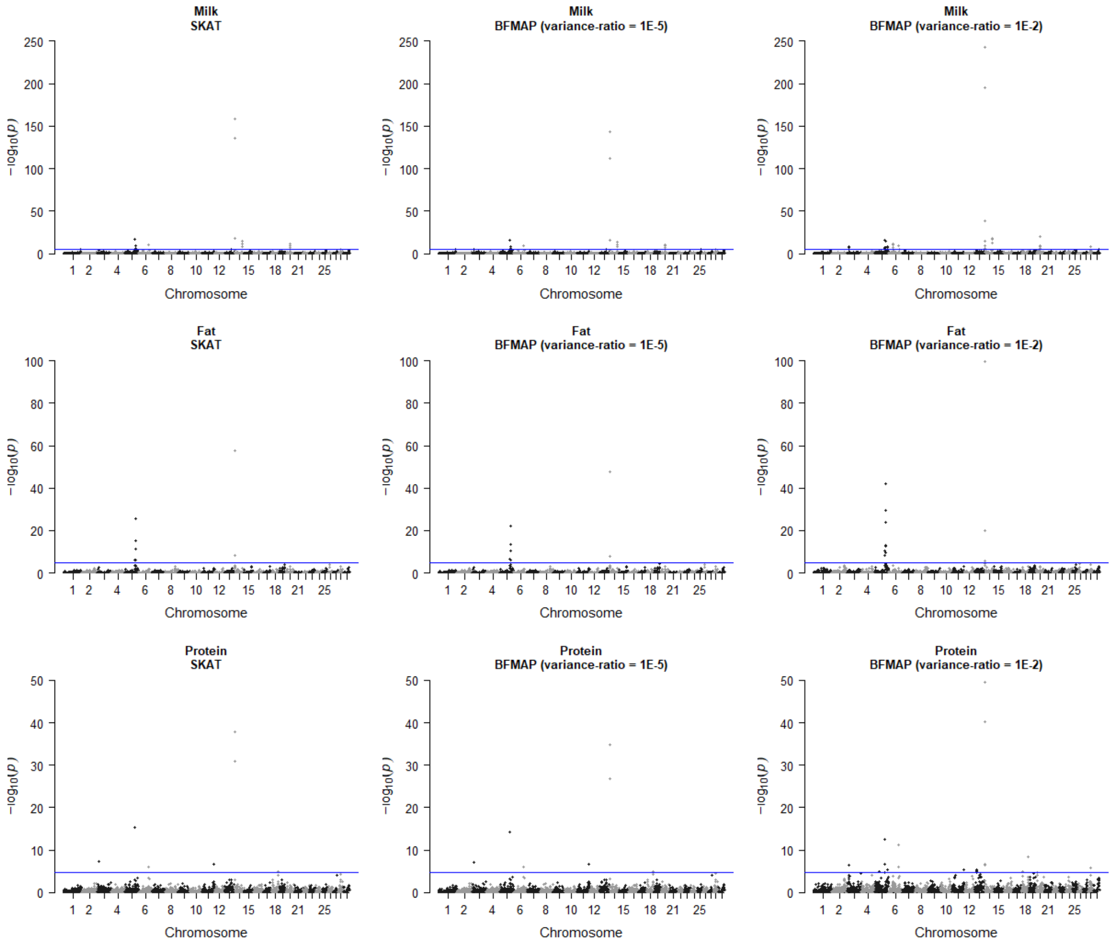


Figure 3.6. SNP-set tests for dairy milk, fat, and protein with BFMAP and SKAT. Variance-ratio means γ in model (3.1).

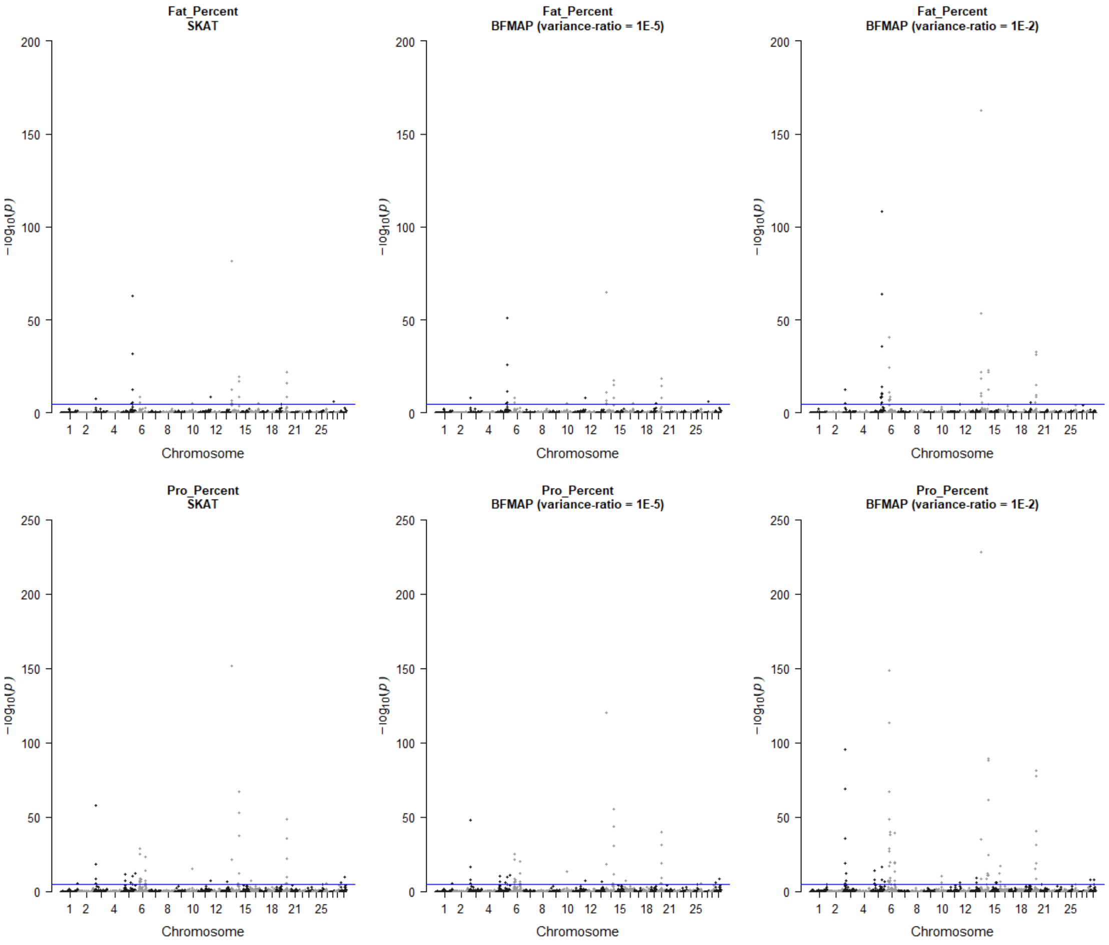


Figure 3.7. SNP-set tests for milk fat percentage (Fat_Percent), and milk protein percentage (Pro_Percent) with BFMAP and SKAT. Variance-ratio means γ in model (3.1).

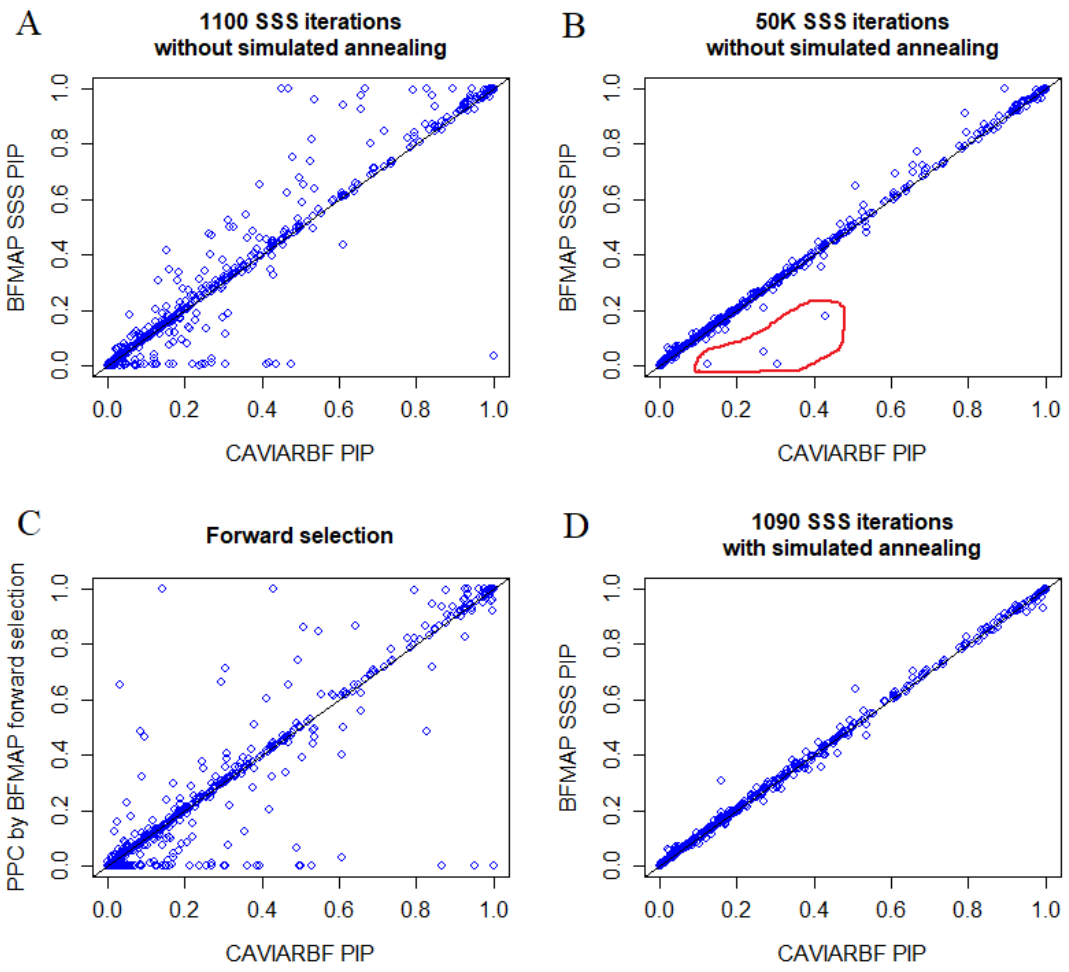


Figure 3.8. Comparisons between BFMAP and CAVIARBF in terms of PIPs of true causal variants using 308 simulated data sets. BFMAP was run with a few different settings. A) BFMAP ran 1100 SSS iterations and simulated annealing was not used. B) BFMAP ran 50 thousand iterations and simulated annealing was not used. C) BFMAP used forward selection procedure. D) BFMAP ran 1090 SSS iterations and default simulated annealing scheme was used (as described in Methods).

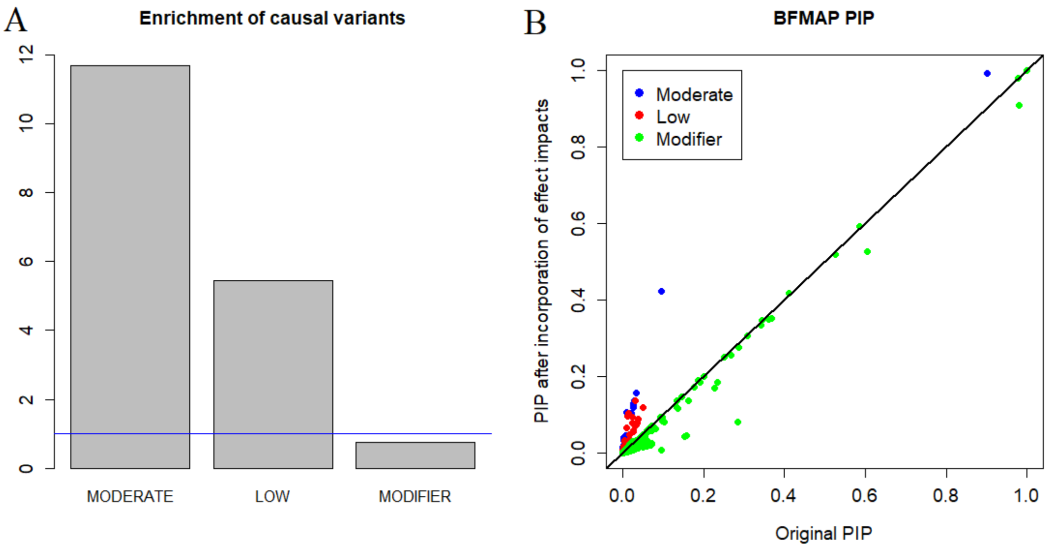


Figure 3.9. Incorporation of SnpEff-inferred variant impact into fine-mapping by BFMAP. A) Estimates of enrichment of causal variants for each category. The blue line is *y*=1. B) PIP changes of variants of each category after incorporation of the functional annotation. The black line is *y*=*x*.

Chapter 4: Incorporating Functional Annotation into Fine-Mapping of 35 Production, Reproduction and Conformation Traits with Imputed Sequences of 27K Holstein Bulls

Abstract

Imputation has been routinely applied to ascertain sequence variants in large genotyped populations based on reference populations of sequenced animals. With the implementation of the 1000 Bull Genomes Project and increasing numbers of animals sequenced, fine-mapping of causal variants is becoming feasible for complex traits in cattle. Using the 1000 Bull Genomes data, we imputed 3 million selected sequence variants to 27,000 Holstein bulls after quality control edits and LD pruning. These bulls were selected to have highly reliable breeding values (PTAs) for 35 production, reproduction, and body conformation traits. We first performed whole-genome single-marker scan for the 35 traits using the mixed-model based association tests. The single-trait association statistics were then merged in multi-trait analyses of 3 groups of traits, production, reproduction, and body conformation, separately. Candidate genomic regions 2 Mb long, were selected based on the multi-trait analyses and used in fine-mapping studies. We used BFMAP to fine-map the dairy cattle traits to single-gene resolution and to integrate fine-mapping with functional enrichment analysis. Our fine-mapping identified many promising candidate genes, including some previously reported ones, e.g., *ABCG2* for production traits and *ARRDC3* for reproduction and body conformation traits. We also show causal effect enrichment patterns for a few functional annotations available in dairy cattle genome and demonstrate that our fine-

mapping result can be readily used for future functional studies. Our study may facilitate follow-up functional validation and expand our understanding of complex traits in dairy cattle. Additionally, our method can be readily applied to other species where large-scale sequence genotypes are available.

Introduction

Phenotypic records have been routinely collected in dairy cattle for over a hundred years. The phenotype of a bull is highly accurately calculated from thousands of phenotypic records from his daughters and other relatives. A comprehensive spectrum of phenotypes has been measured in the cattle population, including production, reproduction, health, and body type traits. GWAS on these traits simultaneously can provide a better understanding of the effects of underlying QTLs. Because of the intensive use of artificial insemination and strong selection in dairy bulls, there are a much smaller number of males than females in the cattle population (Brotherstone and Goddard, 2005), so a chromosome segments can be quickly traced back to an ancestral bull. This uniquely high relatedness in the cattle population can provide accurate imputation (van Binsbergen *et al*, 2014), especially with the reference genomes of important ancestor bulls sequenced by the 1000 Bull Genomes project (VanRaden *et al*, 2017).

Fine-mapping has been commonly performed in human GWAS studies, e.g., (Farh *et al*, 2015; Huang *et al*, 2017). Because of the high linkage disequilibrium levels in the cattle population (Kim and Kirkpatrick, 2009), fine-mapping of GWAS signals has been difficult. In our study, the large sample size can provide enough power to fine-map the major GWAS signals at least to the candidate gene level. The fine-mapped candidate genes will be useful for future functional studies, including the FAANG and related projects in cattle (Andersson *et al*, 2015).

Biologically meaningful enrichment of functional annotation data has been reported in human GWAS (Finucane *et al*, 2015; Sveinbjornsson *et al*, 2016). The high

LD in cattle makes such enrichment difficult to show up in cattle GWAS. With the large sample size and superior power of our study, we hope to identify biologically informative enrichment of variants in our GWAS and fine-mapping results, which can provide useful prior information for future cattle GWAS and genomic selection/prediction.

Results

We imputed three million selected sequence variants to 27,214 Holstein bulls after quality control edits, using the 1000 Bull Genomes data as reference. These bulls were selected to have highly reliable breeding values (PTA) for 35 production, reproduction, and body conformation traits, with an average reliability of 0.71 across traits (Table 4.1). The numbers of bulls available for individual traits ranged from 11,713 to 27,161, with >20,000 animals for 32 traits (Table 4.1). This large, high-quality data set enables our following GWAS and fine-mapping studies with great power and precision.

Single-trait GWAS

We used the mixed model approach implemented in MMAP for single-trait GWAS that can incorporate reliability variation across individual bulls. The mixed model used in our GWAS was robust against potential confounding factors. As shown in Table 4.2, 27 out of the 35 traits had a genomic control factor between 0.95 and 1.05.

We found many clear association signals for the 35 dairy traits. There were in total 286 associations identified for the 35 traits, and the number of associations for individual traits ranged from <3 for leg and foot traits to 23 for protein percentage (Table 4.2). As compared to the Cattle QTLdb release 35, we found that 123 associations (43.0%) had been previously reported while 163 associations (57.0%) were newly discovered in this study. We identified 15 new association signals (out of 68) even for five production traits that had been well studied, while 92 new associations (out of 125) for type traits that drew less attention than other traits in previous studies (Fig. 4.1). The result demonstrated an unprecedented power of our single-trait GWAS in dairy cattle.

Multi-trait association analysis

Hierarchical clustering based on absolute correlation coefficients was largely consistent with the trait definitions: the 35 dairy traits were grouped into three clusters, including production, reproduction, and body type (Fig. 4.2). Interestingly, rump angle and teat length were clustered into reproduction traits, although they are type traits by definition, indicating a close genetic correlation between these two traits and dairy reproduction.

In the multi-trait association analyses for the three trait clusters, we identified 33, 21 and 39 associations for production, reproduction, and type traits using P < 5E-8, respectively (Fig. 4.3). Though a majority of the multi-trait associations were consistent with single-trait ones, we identified ten associations that were missed by single-trait analyses. Based on the multi-trait results, we found two features of multi-trait association tests. First, multi-trait GWAS was more powerful than individual single-trait analyses for related traits. Second, the top variant in multi-trait analysis may be >1 Mb away from the top variants in single-trait GWAS.

Fine-mapping

Initially, we fine-mapped 434 association signals for 282 QTLs applying a significance threshold of 5E-7. The observed distribution of number of fine-mapped signals in a QTL is approximately exponential, which is consistent with our expectation of observing more causal mutations at a QTL with a lower probability (Fig. 4.4). After further quality control edits, we finally determined 308 association signals for 32 traits (Table 4.3). Specifically, there were ≥20 independent association signals identified on chromosomes 5, 6, 14, 18, and 29, while only one or none identified on chromosomes 12, 22, and 27.

Our method enables easy incorporation of functional annotation in fine-mapping. We investigated impacts of incorporation of SnpEff-inferred effect impact (one of the most commonly used functional annotations) on fine-mapping performance. First, we found that incorporating variant impact resulted in substantial change of posterior probability of being causal (PPC) for variants in the fine-mapped 308 association signals. Variants with moderate impact had a considerable increase in PPC when integrating PPC calculation with variant impacts, while modifier variants generally had a decreased PPC (Fig. 4.5A). Second, fine-mapping by incorporating variant impact generated significantly smaller 95% credible variant sets than that using an equal prior for variants, as demonstrated by a Wilcoxon signed-rank test on the 308 signals (*P*<0.01) and Fig. 4.5B. These two features make incorporation of functional annotation favored in practice of fine-mapping.

Enrichment analysis

We first categorized variants into five groups based on their locations regarding protein-coding genes, i.e., CDS, 5’ UTR + 2 kb upstream, intron, 3’ UTR + 2 kb downstream, and other (intergenic or non-protein-coding genic regions). Despite the strong linkage disequilibrium levels in the cattle genome (Bohmanova *et al*, 2010), we observed distinctive enrichment patterns across these five categories (Fig. 4.6A). Using bootstrapping, we calculated 95% confidence intervals for the enrichment levels (Fig. 4.6A), showing significant enrichment of causal variants in CDS (4.52x) and 5’ UTR (2.39x), but not in intron (0.93x) or 3’ UTR (0.77x). We also analyzed a group of non-protein-coding genes and found a serious depletion in this category, $\widehat{E_C}$ = 3.23E-04, suggesting an insignificant effect on the dairy cattle traits.

We further investigated the enrichment of causal variants regarding their genomic locations and protein coding effects (High, Moderate, Low or Modifier) predicted by SnpEff (Cingolani *et al*, 2012). When modeling these four categories, we found a severe depletion of variants with high impact, $\widehat{E_C}$ = 2.51E-05. This is strikingly different from a previous study on human complex traits and diseases that reported an enrichment of >100 for this category (Sveinbjornsson *et al*, 2016). As shown in Fig 5B, we observed a significant enrichment in moderate-impact variants ($\widehat{E_C}$ = 8.7; P < 0.05). Low-impact variants also showed an enrichment ($\widehat{E_C}$ = 2.0), though it was not statistically significant (Fig. 4.6B). As expected, a small depletion was seen in modifier variants (0.87x).

We also used constrained elements on cattle genome to categorize variants into two groups (inside or outside constrained elements), as highly conserved DNA sequences may imply functional importance. As shown in Fig. 4.6C, causal variants were significantly enriched in constrained elements (3.72x; P < 0.05). When further categorizing variants into six groups based on both constrained elements and variant impacts (Moderate, Low or Modifier), we found the highest enrichment in moderate-impact variants inside constrained elements (25.56x; P < 0.05). For other categories, we did not observe significant enrichment of causal variants (Fig. 4.6D).

When comparing different trait groups, we observed little difference in the patterns of enrichment regarding SnpEff-inferred effect impact (Fig. 4.7). Moderate-impact variants had a clearly higher enrichment of being causal for production traits than for reproduction and type traits. We further used permutation to generate the null distribution of E_C(Production)/E_C(Reproduction+Type) and showed that the difference was statistically significant ($P < 0.05$). However, the enrichment for low-impact variants was similar between the three trait groups.

Candidate genes

Based on PPCs of variants after incorporation of SnpEff impact, we calculated PPC for each gene in each independent association signal. There were a total of 564 gene-trait association pairs with PPC >0.01. Most of the genes had either a big PPC (>0.95) or a small one (<0.05). We further obtained a short list of most promising candidates by applying the following conservative criteria: PPC >0.9 if a gene affects only one trait and PPC >0.5 for all traits if a gene affects multiple traits.

This short list had 69 unique genes including some previously reported ones (Table 4.4). For example, *ABCG2* and *DGAT1* are well-known to affect milk production in dairy cattle (Cohen-Zinder *et al*, 2005; Grisart *et al*, 2004). The *ARRDC3* gene has been associated with body confirmation traits and calving traits in beef cattle (Bolormaa *et al*, 2014; Saatchi *et al*, 2014) and Holstein cattle (Abo-Ismail *et al*, 2017). Our fine-mapping study also revealed novel gene/association combinations for dairy traits. A previous study reported that the *ABCC9* gene was associated with fat yield, protein yield and calving to first service interval in Holstein cattle (Nayeri *et al*, 2016). In our study, we discovered that it had a pleiotropic effect on type traits (fore udder attachment and udder depth), milk production (milk and protein yield) and daughter pregnancy rate, with a PPC of almost 1 for all the traits. In addition, we found that there were no common variants among the credible variant sets for these traits (Table 4.4), suggesting that *ABCC9* might have multiple causal mutations for the associated traits. *TMTC2* has been associated with teat length (Abo-Ismail *et al*, 2017), while our fine-mapping showed that it has an effect on six type traits including teat length, with PPC being ≥0.95 for all those traits. Abo-Ismail et al. reported *CCND2* was associated with stature (Abo-Ismail *et al*, 2017). Our fine-mapping determined that it is a candidate gene for four type traits (PPC >0.95 for body depth, rump width and stature). It is worth noting that our fine-mapping study not only discovered association of a gene with a trait, but also provided posterior probability of being causal for a gene.

Candidate variants

Considering that our stringent QC during and after imputation removed many variants, fine-mapping the traits to single-variant resolution could not always be achieved. Nevertheless, we obtained 95% credible variant set for each independent signal and merged them into one table. This resulted in a total of 1,582 unique variants. We generated a short list by keeping only variants with moderate impact and PPC >0.1 (Table 4.5). Among the list, some variants have been well studied, e.g., Chr6:38027010 in *ABCG2* (Cohen-Zinder *et al*, 2005) and Chr26:21144708 in *SCD* (Pegolo *et al*, 2016). We also found other promising candidate variants, e.g., Chr8:83581466 in *PTH1* with an average PPC of 0.68 on two genetically correlated type traits (body depth and strength), Chr1:69673871 in *KALRN* with an average PPC of 0.46 on two genetically correlated fertility traits (cow conception rate and daughter pregnancy rate), Chr17:70276788 in *CHEK2* with an average PPC of 0.39 on two highly correlated calving traits (sire calving ease and daughter calving ease).

Discussion

In this study, we performed GWAS for 35 production, reproduction, and type traits in dairy cattle with a unique large-scale data set, and further fine-mapped these traits to single-gene resolution. With the fast computing tool that we developed, we attempted to find causal effects in hundreds of loci each of which contains thousands of variants. We also investigated the functional enrichment patterns of several functional annotations available in dairy cattle, and incorporated the information into fine-mapping. By the study, we provide not only a credible candidate gene list for follow-up functional validation, but also a unique resource that can be easily used by future functional studies.

Single-trait GWAS

In the single-trait GWAS, we find many association signals that have not been discovered (Fig. 4.1), clearly demonstrating the benefits of using the unique large-scale dairy cattle data. Reliabilities of de-regressed PTAs were modeled for most of the traits (Table 4.2). For the traits with small variation of reliability, we observed similar results for the models with and without reliability; e.g., QTLs found when not modeling reliability were largely the same as those by incorporating reliability for fat percentage and daughter pregnancy rate. Interestingly, we observed some deflations in GWAS of production traits, which could be due to the large QTL effects on these traits including the *DGAT1* gene. Minor inflations were observed in GWAS for calving traits (i.e., calving ease and stillbirth) and final score. Although there were sporadic variants passing the threshold of genome-wide significance (P < 5E-8), we could locate a few GWAS peaks where there were a cluster of significant variants.

Multiple testing in fine-mapping

Initially, our fine-mapping discovered as many as 19 signals in a candidate region for a trait, as it applied a variant inclusion threshold accounting for only the effective number of independent variants (m_{eff}) at locus-by-trait level. We also noticed that there were more locus-by-trait association pairs with multiple signals than with one signal. By examining those with multiple signals, we found the models often contained a strong signal and much weaker one(s). Those weak signals might result from imperfect model fitting of lead variants in other signals, instead of being true positive. Nevertheless, they did little harm to the discovery of true signals.

Enrichment of causal variants

The enrichment estimates for SnpEff-inferred variant impact in our study are very different from those in a previous human study (Sveinbjornsson *et al,* 2016). The differences among the four categories in the human study are much more distinctive than ours. This is consistent with our anticipation that the high LD in cattle makes such enrichment difficult to show up. Nevertheless, we find a considerable enrichment of causal effects in moderate-impact variants. Incorporation of the enrichment into fine-mapping facilitates the discovery of causal variants (Fig. 4.5). The discovery of the enrichment patterns is also valuable for development of functional annotation-driven methods for better genomic prediction.

Fine-mapping

By fine-mapping, we pinpoint some promising candidate genes for economically important traits in dairy cattle. It is promising to validate those genes with high posterior probability of causality. In addition, with our method of functional enrichment analysis, our fine-mapping result of hundreds of QTLs (basically, variant PPCs) can be readily used for functional annotations other than those analyzed here. Thus, we provide an easy-to-use enrichment analysis procedure to analyze the functional annotations that the FAANG and related projects will produce on cattle genome.

Materials and Methods

Genotype and phenotype data

Genotype data have been described in detail in our previous study (VanRaden *et al*, 2017). Here we give a summary. SNP and insertion-deletion (InDel) calls (sequence variants) from run 5 of the 1000 Bull Genomes Project (Daetwyler *et al*, 2014) were released in July 2015. After stringent quality control edits (without LD pruning), 3,148,506 sequence variants remained for 444 Holstein animals. The sequence variant data and high-density (HD) genotypes of 312K markers for 26,949 progeny-tested Holstein bulls (and 21 Holstein cows) were combined by imputation using findhap software (version 3) (VanRaden, 2016b). Finally, we had (imputed) genotypes of 3,148,506 sequence variants for 27,214 Holstein bulls (179 bulls had both sequence and HD genotypes) and 21 cows.

Imputation quality from findhap software was assessed with 404 of the sequenced animals in the reference population and 40 randomly chosen animals for validation. Their sequence genotypes were reduced to the subset of genotypes that were in common with HD genotypes and then imputed back to sequence. Imputation accuracy was equal to 96.7% for the 3,148,506 variants (VanRaden *et al*, 2017). If HD

SNPs were not counted, we found an accuracy of 96.4% for just the new variants. Chromosome-specific imputation accuracy was >95% for all autosomes except chromosome 12.

All the 27,214 Holstein bulls were selected to have highly reliable predicted transmitting abilities (PTAs) for 35 production, reproduction, and type traits, although not all bulls had PTAs for all the traits. Transmitting ability is basically additive genetic value accounting for additive genetic variance. Reliability quantifies the amount of information available in a PTA and measures its accuracy (VanRaden and Wiggans, 1991). De-regressed PTAs were used as phenotypes in all our analyses, which excludes parent information and reduces dependence in PTAs among animals (Garrick *et al*, 2009). Because each of the bulls usually had many phenotyped daughters that were used for breeding value estimation, their PTAs were generally of high reliability, even for low-heritability reproduction traits (Table 4.1). We can largely categorize the traits into three groups, i.e. production, reproduction and type.

Single-trait GWAS

The software MMAP (O'Connell, 2013) was used for all single-trait GWAS analyses. Basically, MMAP efficiently implements a mixed-model approach for association tests which is similar to GEMMA (Zhou and Stephens, 2012) but different from EMMAX (Kang *et al*, 2010); that is, variance component is estimated uniquely for each marker. We used the following model

$$\mathbf{y} = \mu + \mathbf{X}b + \mathbf{g} + \mathbf{e} \text{ with } \mathbf{g} \sim N\left(0, \sigma_g^2 \mathbf{G}\right) \text{ and } \mathbf{e} \sim N(\sigma_e^2 \mathbf{R}), \qquad (4.1)$$

where $\boldsymbol{y}$ is de-regressed PTAs, μ is global mean, $\boldsymbol{X}$ is genotype of a variant (coded as 0, 1 or 2) and b is its effect, $\boldsymbol{g}$ is polygenic effect accounting for population structure, and

e is residual. The genomic relationship matrix (***G***) (VanRaden, 2008) was built using 312K HD markers (filtered by MAF>1%). ***R*** is a diagonal matrix ($\boldsymbol{R}_{ii} = \mathbf{1}/\boldsymbol{r}^{2} - \mathbf{1}$), which is used to model differential reliability among animals.

We disregarded variants on the X chromosome. We also filtered out variants with an MAF of <1% or failing Hardy-Weinberg equilibrium (HWE) test (p < 1E-6). After the QC, there were ~2.7 million variants left. QTLs were located by finding GWAS peaks where there were a cluster of significant variants. We used a custom Perl script to find all GWAS peaks and further examined each of the peaks based on Manhattan plots to keep only clear ones. Subsequently, we determined a total of 286 QTLs which were further analyzed in fine-mapping studies.

To find which ones are novel among the 286 QTLs, we compared our result with Cattle QTLdb (release 35 published on April 29, 2018) which contains 113,256 QTLs/associations from 848 publications (Hu *et al*, 2016). To ensure correct physical position of QTLs/associations on UMD 3.1, we first extracted rs identifiers (rs#) of flanking markers for each term from the Cattle QTLdb data, and then used the identifiers to find flanking markers' positions on UMD 3.1 in the Ensembl genome variation database. These marker positions were used as QTL/association positions. This procedure can rule out QTL terms whose physical positions are inaccurately converted from genetic map. The Cattle QTLdb release 35 covers 599 different traits, in which we found the ones with the (almost) same definition as our 35 traits. For each of the QTLs that we detected, we determined that it had been previously reported if it is within ±500 kb of any QTL/association for the (almost) same trait(s) in Cattle QTLdb and that it was newly discovered otherwise.

Multi-trait association analysis

Following a previous study (Bolormaa *et al*, 2014), our multi-trait association tests were based on a chi-square statistic with multiple degrees of freedom. For each variant, the chi-square statistic for the multi-trait association test was calculated by the formula:

$$\text{Multi-trait } \chi^2\left(d.f.=n\right)=\mathbf{t}_i'\mathbf{V}^{-1}\mathbf{t}_i\ ,$$

where $\boldsymbol{t}_i$ is a $n \times 1$ vector of the signed t-values of variant *i* for *n* traits, and $\boldsymbol{V}$ is an $n \times n$ correlation matrix for the *n* traits which is calculated using signed t-values of genome-wide variants. In our analysis, the signed t-values were obtained from single-trait GWAS for 2,619,418 variants passing QC, and the correlations between traits were calculated using all the variants.

To test the robustness of the estimated correlation using all sequence variants (Zhu *et al*, 2015a), we also computed the correlation matrix using two variant subsets obtained by selecting every 10th and every 100th variant. The three variant sets produced similar correlation estimations.

We performed hierarchical clustering based on absolute correlation coefficients, and then did multi-trait association analysis for each of the three resulting clusters of traits as shown in Fig. 4.2. Specifically, we excluded net merit and DFB in production and reproduction, respectively, since both the traits are basically linear combinations of other traits (and the number of bulls for DFB was much smaller than those for other traits). We also excluded the four calving traits to avoid the contamination by sporadic significant variants. Additionally, all the traits except for the six traits aforementioned were analyzed as a whole in multi-trait association tests.

We identified ten associations in multi-trait analyses that were missed in single-trait analyses. Some individual traits showed suggestive association (P < 5E-6) in these ten loci, which were added to the following fine-mapping studies.

Bayesian fine-mapping approach

Our Bayesian approach for fine-mapping has been well described in Chapter 3. In this chapter, we focus on the use of forward selection approach in BFMAP, especially how to integrate forward selection results with functional annotation. To make this chapter easier to read, the model is described again. Note that the model used here is a simplified version of model (3.1), in that the diagonal matrix for variant weights is replaced by an identity matrix, shown as follow:

$$\begin{gathered}\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{g} + \mathbf{e} \\ \mathbf{b} \sim N(0, \varphi\sigma_e^2\mathbf{I}) \\ \mathbf{a} \sim N(0, \gamma\sigma_e^2\mathbf{I}) \\ \mathbf{g} \sim N(0, \eta\sigma_e^2\mathbf{G}) \\ \mathbf{e} \sim N(0, \sigma_e^2\mathbf{R}) \\ P\left(\sigma_e^2\right) \propto 1/\sigma_e^2\end{gathered}, \qquad (4.2)$$

where $\boldsymbol{y}$ is a phenotype vector of size n for a complex trait, $\boldsymbol{b}$ is a vector of covariate (other than genomic variants) effect and $\boldsymbol{X}$ is corresponding design matrix, $\boldsymbol{a}$ is a vector of variant effect and $\boldsymbol{Z}$ is corresponding genotype coding matrix (e.g., genotype coding for additive, dominance or imprinting effects (Jiang *et al*, 2017)), $\boldsymbol{g}$ is a vector of polygenic effect for controlling population structure and $\boldsymbol{G}$ is corresponding variance structure matrix (e.g., genomic relationship matrix), and $\boldsymbol{e}$ is residual with variance structure $\boldsymbol{R}$ for modelling reliability or accuracy of phenotypic records as in model (4.1). The common variance component (σ_e^2) is given a non-informative Jeffrey's prior. Other variance parameters (φ, γ and η) are treated as known. Generally, we can set φ to a large value (e.g., 1E8) to make $\boldsymbol{a}$ act like fixed effects. A genomic variant is usually considered to be of small but noticeable effect, so we can set γ to 0.01 or 0.04 (Chen *et al*, 2015; Zhou and Guan, 2017a). When $\boldsymbol{Za}$ only accounts for a tiny proportion of phenotypic variance (this is true when modeling variants from a small genomic region), we can set η based on heritability (h^2), $\eta = h^2/(1-h^2)$. In practice, we can instead use heritability estimate ($\widehat{h^2}$) in the null model without variants to determine η. In the context of GWAS, we are only interested in variant effects ($\boldsymbol{a}$).

We aim to identify independent association signals within a region and to assign a posterior probability of causality (PPC) to each variant with fine-mapping. Following the first method of (Huang *et al*, 2017), our fine-mapping approach includes three steps: forward selection (Foster and George, 1994) to add independent signals in the model, repositioning signals, and generating credible variant set for each signal.

We set $\varphi = \gamma = 1E8$ in model (4.2) for fine-mapping, which enables easy calculation of *p*-value for a newly added variant conditioning on variants being already in model. We use Bonferroni threshold (Foster and George, 1994) as stopping criterion in forward selection; that is, forward selection stops when $\left(2\log \mathrm{sBF}+1\right) < 2\log m_{\mathrm{eff}}$, where m_{eff} is efficient number of independent variants calculated using the method by Li and Ji (2005). Suppose that we select *p* independent signals in forward selection and determine a set of lead variants (S_l) for the *p* signals after repositioning. Then for signal *i* with lead variant (l_{i}), we have a variant set (S_i) containing variants that have substantial LD with l_{i} but weak LD with lead variants in other signals $S_l \setminus \{l_i\}$. Accordingly, we can compute PPC of variant *j* (v_{ij}) in S_i conditioning on $S_l \setminus \{l_i\}$:

$$P(M_i = v_{ij} \mid y, X, Z, S_l \setminus \{l_i\}) = \frac{P(y \mid X, Z, M_i = v_{ij}, S_l \setminus \{l_i\}) P(M_i = v_{ij})}{\sum_j P(y \mid X, Z, M_i = v_{ij}, S_l \setminus \{l_i\}) P(M_i = v_{ij})}, \quad (4.3)$$

where $M_i = v_{ij}$ denotes that the causal variant in signal i is variant j in S_i (i.e. v_{ij}). Efficient computation of $P(y \mid X, Z, M)$ has been described in Chapter 3. We can easily get a credible variant set passing a given confidence level (e.g., 95%) for a signal, by sorting variants in a descending order of PPC and including them in the set from top. We can also calculate PPC of a gene by summing up PPCs of all variants within the gene.

In the study by Huang *et al* (2017), an equal prior for each variant was used; that is, $P(M_i = v_{ij}) = 1 \; \forall v_{ij} \in S_i$. Here we propose a method to apply differential prior probabilities by integrating functional annotation, drawing ideas from a previous study on adjusting significance threshold based on functional annotation in GWAS (Sveinbjornsson *et al*, 2016). With our fine-mapping procedure, it is usually safe to assume that there is only one causal variant in each independent signal. For a function annotation with several categories, we denote the probability of a causal variant being of category C as p_C and the probability of a non-causal variant being of category C as q_C. We can accordingly obtain:

$$P(M_i = v_{ij}) = P(c_{ij} \mid M_i = v_{ij}) \prod_{j' \neq j} P(c_{ij'} \mid M_i \neq v_{ij'}) = p_{c_{ij}} \prod_{j' \neq j} q_{c_{ij'}} , \quad (4.4)$$

where c_{ij} denotes the category of variant j in S_i (i.e. v_{ij}).

We estimate q_C with the genome-wide frequencies of the categories, as in (Sveinbjornsson *et al*, 2016). To estimate p_C, we can use all the available independent signals (M_i):

$$\begin{aligned}&L(\{p_C\}\mid y,Z)\\&\propto\prod_i P(M_i,y,Z\mid\{p_C,q_C\})\\&\propto\prod_i\sum_j P(y\mid X,Z,M_i=v_{ij})P(M_i=v_{ij}\mid\{p_C,q_C\})\end{aligned}\quad. \tag{4.5}$$

When the signals identified in fine-mapping are independent of each other, which is generally true with our approach, we can get:

$$P(y\mid X,Z,M_i=v_{ij})\doteq P(y\mid X,Z,M_i=v_{ij},S_l\setminus\{l_i\})\,. \tag{4.6}$$

Taking equations (4.4) and (4.6) into equation (4.5), we obtain a likelihood function regarding $\{p_C\}$ and then get their maximum likelihood estimates (MLEs), $\{\hat{p}_C\}$. By taking the estimates of $\{p_C,q_C\}$ and equation (4.4) to equation (4.3), we get updated PPCs with incorporation of function annotation, which is actually an empirical Bayes approach.

When setting an equal prior for each variant, we find:

$$P(M_i=v_{ij}\mid y,X,Z,S_l\setminus\{l_i\})\propto P(y\mid X,Z,M_i=v_{ij},S_l\setminus\{l_i\})\,. \tag{4.7}$$

Thus, to estimate $\{p_C\}$ by equation (4.5), we can use PPCs from the computation assuming an equal prior for each variant. Accordingly, incorporation of functional annotation includes three separate steps: computing PPCs given an equal prior for each variant, estimating $\{q_C\}$ with the genome-wide frequencies of the categories and estimating $\{p_C\}$ with these PPCs, and updating PPCs with $\{\hat{p}_C,\hat{q}_C\}$. This feature makes our approach easier to use compared with PAINTOR (Kichaev *et al*, 2014) and CAVIARBF (Chen *et al*, 2016).

Fine-mapping dairy cattle traits

Genomic regions for find-mapping were determined by lead variants in single-trait and multi-trait QTLs. Lead variants in a candidate genomic region may be different between multi-trait QTL and single-trait QTLs. Accordingly, we first determined a minimal region that covered all the lead variants (either in multi-trait or in single-trait QTLs), and then extended it 1 Mb upstream and downstream, which resulted in a ≥2 Mb genomic region used for fine-mapping. The 1-Mb extensions allowed the region to cover almost all variants that have an LD r^2 of >0.3 with lead variants (Bohmanova *et al*, 2010).

Subsequently, we obtained a total of 125 loci. Three loci without plentiful HD SNP markers were removed to ensure imputation quality, thus leaving 122 loci in fine-mapping. Fifty-seven loci were associated with more than one trait. The fine-mapping was performed for individual traits, and these 122 loci represented 282 locus-by-trait association pairs for 32 traits (three leg type traits were excluded for lack of significance). When fine-mapping identified multiple signals in a candidate locus for a trait, we kept the strongest one and filtered the rest. The effective number of independent tests was 54,403 for the 282 locus-by-trait pairs. Considering that our effective number estimates were conservative (Hendricks *et al*, 2014), we used 5E-7 (<0.05/54,403) as the significance threshold to filter signals. Subsequently, we found 434 association signals.

We found that the locus-by-trait association pairs with more than three signals identified were mostly from still birth and final score. We also noticed slight inflation of the test statistics in the GWAS of these traits. Therefore, we removed the 16 QTLs with >3 fine-mapped signals in our following analyses. We further removed 15 signals whose variant set had ≤10 variants of distinct genotypes, as a small cluster of highly linked variants could be due to inaccurate imputation. Additionally, if there were multiple QTL on a chromosome for a trait, all lead variants in these loci were modeled jointly in fine-mapping. Accordingly, 13 association signals whose lead variant had a *p*-value of >5e-7 were removed. After all the edits, we determined a total of 308 association signals (Table 4.3).

Besides assuming an equal prior for each variant, we further applied differential prior probabilities based on SnpEff-inferred effect impacts (Cingolani *et al*, 2012). Since using equation (4.5) requires independent association signals, we removed all association signals for protein, cow conception rate, rear teat placement, udder depth and strength, because they have high correlation (r^2>0.5) with other traits. We also removed another six association signals, since these signals have a substantial LD with another signal (measured by LD r^2 between lead variants >0.25). These edits reduced the number of association signals from 308 to 249. We estimated $\{p_C, q_C\}$ for variant impact categories based on the 249 association signals, and updated PPCs for all 308 signals by integrating the estimates.

Effect impact-incorporated PPCs were used for determining candidate mutations or genes. When computing PPC of a gene, variants within its two-kb upstream/downstream were included besides those within the gene.

Enrichment analysis

Our enrichment analysis was based on our fine-mapped 249 association signals (as described above) to estimate p_C (the probability of a causal variant being in category

C) and q_C (the probability of a non-causal variant being in category *C*). The enrichment for category *C* is defined as $E_C = p_C/q_C$ (Sveinbjornsson *et al*, 2016), for which a value larger than one indicates that causal variants are more enriched in category *C* than across whole genome. Functional annotations investigated included locations of variants regarding protein-coding genes, effect impact inferred by SnpEff (Cingolani *et al*, 2012), and constrained elements predicted by GERP (Cooper *et al*, 2005). Confidence intervals of the enrichment estimates was derived by percentile bootstrap as in (Sveinbjornsson *et al*, 2016). The association signals were sampled 1,000 times to calculate each confidence interval. We removed very small categories (like HIGH in SnpEff-inferred effect impacts) in bootstrapping, since including them often resulted in bad convergence of maximum likelihood estimation.

URLs

BFMAP: http://terpconnect.umd.edu/~jiang18/bfmap/

MMAP: https://mmap.github.io/

Cattle constrained elements: ftp://ftp.ensembl.org/pub/release-90/bed/ensembl-compara/68_eutherian_mammals_gerp_constrained_elements/gerp_constrained_elements.bos_taurus.bed.gz

Cattle genome annotation:

ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000003055.6_Bos_taurus_UMD_3.1.1/GCF_000003055.6_Bos_taurus_UMD_3.1.1_genomic.gff.gz

Cattle QTLdb: https://www.animalgenome.org/cgi-bin/QTLdb/BT/index

Cattle genome variation: ftp://ftp.ensembl.org/pub/release-89/variation/gvf/bos_taurus/

References

Abo-Ismail MK, Brito LF, Miller SP, Sargolzaei M, Grossi DA, Moore SS *et al* (2017). Genome-wide association studies and genomic prediction of breeding values for calving performance and body conformation traits in Holstein cattle. *Genet Sel Evol* **49**(1)**:** 82.

Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW *et al* (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol* **16:** 57.

Bohmanova J, Sargolzaei M, Schenkel FS (2010). Characteristics of linkage disequilibrium in North American Holsteins. *BMC Genomics* **11:** 421.

Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K *et al* (2014). A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS Genet* **10**(3)**:** e1004198.

Brotherstone S, Goddard M (2005). Artificial selection and maintenance of genetic variance in the global dairy cow population. *Philos Trans R Soc Lond B Biol Sci* **360**(1459)**:** 1479-1488.

Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA *et al* (2015). Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**(3)**:** 719-736.

Chen W, McDonnell SK, Thibodeau SN, Tillmans LS, Schaid DJ (2016). Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics. *Genetics* **204**(3)**:** 933-958.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L *et al* (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**(2)**:** 80-92.

Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee JH *et al* (2005). Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res* **15**(7)**:** 936-944.

Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S *et al* (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**(7)**:** 901-913.

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF *et al* (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46**(8)**:** 858-865.

Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S *et al* (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**(7539)**:** 337.

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR *et al* (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**(11)**:** 1228-1235.

Foster DP, George EI (1994). The Risk Inflation Criterion for Multiple-Regression. *Ann Stat* **22**(4)**:** 1947-1975.

Garrick DJ, Taylor JF, Fernando RL (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol* **41:** 55.

Grisart B, Farnir F, Karim L, Cambisano N, Kim JJ, Kvasz A *et al* (2004). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc Natl Acad Sci U S A* **101**(8)**:** 2398-2403.

Hendricks AE, Dupuis J, Logue MW, Myers RH, Lunetta KL (2014). Correction for multiple testing in a gene region. *Eur J Hum Genet* **22**(3)**:** 414-418.

Hu ZL, Park CA, Reecy JM (2016). Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Res* **44**(D1)**:** D827-833.

Huang H, Fang M, Jostins L, Umicevic Mirkov M, Boucher G, Anderson CA *et al* (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**(7662)**:** 173-178.

Jiang J, Shen B, O'Connell JR, VanRaden PM, Cole JB, Ma L (2017). Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. *BMC Genomics* **18**(1)**:** 425.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB *et al* (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**(4)**:** 348-354.

Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL *et al* (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**(10)**:** e1004722.

Kim ES, Kirkpatrick BW (2009). Linkage disequilibrium in the North American Holstein population. *Anim Genet* **40**(3)**:** 279-288.

Li J, Ji L (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**(3)**:** 221-227.

Nayeri S, Sargolzaei M, Abo-Ismail MK, May N, Miller SP, Schenkel F *et al* (2016). Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. *BMC Genet* **17**(1)**:** 75.

O'Connell JR. (2013). *63th Annual Meeting of The American Society of Human Genetics*.

Pegolo S, Cecchinato A, Mele M, Conte G, Schiavon S, Bittante G (2016). Effects of candidate gene polymorphisms on the detailed fatty acids profile determined by gas chromatography in bovine milk. *J Dairy Sci* **99**(6)**:** 4558-4573.

Saatchi M, Schnabel RD, Taylor JF, Garrick DJ (2014). Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. *BMC Genomics* **15:** 442.

Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A, Masson G *et al* (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet* **48**(3)**:** 314-317.

van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I *et al* (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol* **46:** 41.

VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**(11)**:** 4414-4423.

VanRaden PM. (2016). *Vol. 2016*: Animal Improvement Program, Animal Genomics and Improvement Laboratory, ARS, USDA. .

VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet Sel Evol* **49**(1)**:** 32.

VanRaden PM, Wiggans GR (1991). Derivation, calculation, and use of national animal model information. *J Dairy Sci* **74**(8)**:** 2737-2746.

Zhou Q, Guan Y (2017). On the Null Distribution of Bayes Factors in Linear Regression. *Journal of the American Statistical Association*(just-accepted).

Zhou X, Stephens M (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**(7)**:** 821-824.

Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N *et al* (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am J Hum Genet* **96**(1)**:** 21-36.

Tables

Table 4.1. Number of Holstein bulls, and mean and standard deviation (SD) of PTAs/reliabilities for each trait

Trait abbreviation	**Direction of selection**	**No. of bulls**	**Deregressed PTAs**		**Reliability of deregressed PTAs**				**Trait name**	**Trait group**
			Mean	SD	Mean	SD	Min	Max		
Milk[1]	+	27156	-245.86	850.58	0.860	0.082	0.325	0.999	Milk yield	Production
Fat[1]	+	27156	-5.92	30.52	0.860	0.082	0.325	0.999	Fat yield	Production
Protein[1]	+	27156	-5.31	23.84	0.863	0.083	0.325	0.999	Protein yield	Production
Fat_Percent[1]	+	27156	0.0136	0.107	0.860	0.082	0.325	0.999	Fat percentage	Production
Pro_Percent[1]	+	27156	0.0086	0.0464	0.863	0.083	0.325	0.999	Protein percentage	Production
Net_Merit	+	27161	-106.91	278.63	0.763	0.110	0.067	0.990	Net merit	
Prod_Life	+	26727	-1.367	3.461	0.682	0.145	0.147	0.999	Productive life	Reproduction
SCS	-	27143	3.027	0.235	0.786	0.110	0.040	0.999	Somatic cell score	
AFC	-	16314	-0.446	11.855	0.439	0.258	0.010	0.990	Age at first calving	Reproduction
DFB[2]	-	11713	0.534	2.825					Days to first breeding	Reproduction
Dtr_Preg_Rate	+	25699	-0.593	3.025	0.618	0.185	0.061	0.999	Daughter pregnancy rate	Reproduction
Heifer_Conc_Rate	+	19334	-0.660	9.610	0.377	0.210	0.002	0.990	Heifer conception rate	Reproduction
Cow_Conc_Rate	+	20380	-1.053	6.879	0.597	0.202	0.002	0.990	Cow conception rate	Reproduction
Sire_Calv_Ease	-	26345	7.959	2.461	0.671	0.224	0.082	0.990	Sire calving ease	Reproduction
Dtr_Calv_Ease	-	23263	9.141	3.182	0.594	0.176	0.160	0.990	Daughter calving ease	Reproduction
Sire_Still_Birth	-	21543	8.190	1.831	0.495	0.249	0.019	0.990	Sire stillbirth	Reproduction
Dtr_Still_Birth	-	20424	8.085	2.958	0.508	0.222	0.040	0.990	Daughter stillbirth	Reproduction
Final_score	+	25638	-0.817	1.484	0.702	0.140	0.144	0.990	Final score	Type

Stature	+	25641	-0.482	1.532	0.844	0.079	0.404	0.990	Stature	Type
Strength	+	25633	-0.278	1.513	0.743	0.147	0.017	0.990	Strength	Type
Dairy_form	null	25615	-0.492	1.745	0.752	0.132	0.149	0.990	Dairy form	Type
Foot_angle	+	25626	-0.742	2.263	0.664	0.198	0.029	0.990	Foot angle	Type
Rear_legs(side)	+	25641	-0.009	1.734	0.754	0.137	0.121	0.990	Rear legs (side view)	Type
Body_depth	+	25636	-0.413	1.622	0.720	0.180	0.060	0.990	Body depth	Type
Rump_angle	null	25641	0.038	1.482	0.828	0.089	0.338	0.990	Rump angle	Type
Rump_width	+	25641	-0.504	1.543	0.766	0.114	0.229	0.990	Rump width	Type
Fore_udder_att	+	25640	-0.908	1.852	0.781	0.112	0.176	0.990	Fore udder attachment	Type
Rear_ud_height	+	25640	-0.885	2.095	0.737	0.136	0.229	0.990	Rear udder height	Type
Udder_depth	+	25631	-0.653	1.665	0.836	0.082	0.355	0.990	Udder depth	Type
Udder_cleft	+	25641	-0.720	1.980	0.718	0.156	0.089	0.990	Udder cleft	Type
Front_teat_pla	+	25641	-0.562	1.663	0.781	0.106	0.324	0.990	Front teat placement	Type
Teat_length	+	25631	0.104	1.482	0.815	0.087	0.355	0.990	Teat length	Type
Rear_legs(rear)	+	24763	-0.759	2.709	0.605	0.178	0.028	0.990	Rear legs (rear view)	Type
Feet_and_legs	+	25608	-0.928	2.501	0.600	0.208	0.027	0.990	Feet and legs composite	Type
Rear_teat_pla	+	25492	-0.436	1.900	0.762	0.103	0.062	0.990	Rear teat placement	Type

[1] Besides bulls, we included in single-trait GWAS two Holstein cows with high reliability (~0.40).

[2] For DFB, we used PTAs instead of deregressed PTAs.

Table 4.2. Genomic control factor of single-trait GWAS for each trait

Trait	Modeling reliability	GC lambda	N.QTLs
Milk	N	0.939	14
Fat	N	0.907	9
Protein	N	1.008	10
Fat_Percent	N	0.753	12
Pro_Percent	N	0.828	23
AFC	Y	1.020	3
DFB	N	1.010	4
Net_Merit	N	1.004	6
Prod_Life	N	0.984	9
SCS	N	0.970	10
Dtr_Preg_Rate	Y	1.022	9
Heifer_Conc_Rate	Y	1.010	3
Cow_Conc_Rate	Y	1.020	7
Sire_Calv_Ease	Y	1.051	8
Dtr_Calv_Ease	Y	1.026	7
Sire_Still_Birth	Y	1.106	8
Dtr_Still_Birth	Y	1.061	7
Final_score	Y	1.054	12
Stature	Y	0.958	13
Strength	Y	0.971	7
Dairy_form	Y	1.022	7
Foot_angle	Y	1.008	2
Rear_legs(side)	Y	1.023	0
Body_depth	Y	0.978	9
Rump_angle	Y	1.016	8
Rump_width	Y	0.967	10
Fore_udder_att	Y	1.019	16
Rear_ud_height	Y	1.034	6
Udder_depth	Y	0.987	14
Udder_cleft	Y	1.020	3
Front_teat_pla	Y	0.980	6
Teat_length	Y	0.963	16
Rear_legs(rear)	Y	1.022	0
Feet_and_legs	Y	1.034	0
Rear_teat_pla	Y	0.974	8

Table 4.3. Number of association signals on each chromosome for each trait

Trait	**Chromosome**																													**Autosome**
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	
Milk	1	0	2	0	3	2	0	0	0	0	0	0	0	3	1	0	0	0	0	1	0	0	0	0	1	0	0	1	1	16
Fat	0	2	0	0	2	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1	0	9
Protein	0	0	1	0	2	2	0	0	0	0	1	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	10
Fat_Percent	0	0	2	0	2	1	0	0	0	0	1	0	0	1	1	1	0	0	1	3	0	0	0	0	0	1	1	0	0	15
Pro_Percent	1	0	2	0	4	4	0	0	1	1	1	0	1	6	2	0	0	1	0	0	1	0	0	0	0	1	0	1	2	29
Net_Merit	0	0	0	0	0	1	1	0	0	0	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	6
Rump_angle	1	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	8
Teat_length	0	0	0	1	5	0	1	0	0	0	2	0	0	0	0	1	0	1	1	1	1	1	0	0	0	2	0	0	1	18
AFC_DYD	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	3
Heifer_Conc_Rate	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3
Dtr_Calv_Ease	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	6
Dtr_Still_Birth	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Sire_Calv_Ease	0	0	0	0	3	0	2	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	8
Sire_Still_Birth	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	4
DFB_PTA	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	4
Dtr_Preg_Rate	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	1	8
Cow_Conc_Rate	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	1	7

Dairy_form	0	0	1	0	2	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	1	0	8
Prod_Life	0	0	0	0	1	1	1	0	0	0	0	0	1	1	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	8
SCS	0	0	0	0	0	1	1	0	0	0	0	0	1	2	1	0	0	1	2	1	0	0	0	0	0	0	0	0	0	10
Udder_cleft	2	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	7
Front_teat_pla	0	0	0	0	2	0	0	2	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	2	0	0	0	9
Rear_teat_pla	1	0	0	0	1	0	2	1	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	1	3	0	0	0	13
Fore_udder_att	0	0	0	1	2	0	1	0	0	0	1	0	2	0	1	0	0	1	0	3	2	0	0	0	0	1	0	1	2	18
Udder_depth	0	0	0	1	1	1	1	1	0	0	2	0	0	1	0	0	0	1	1	2	0	0	1	0	0	0	0	0	2	15
Final_score	0	0	0	0	2	0	0	2	0	0	1	0	0	0	1	0	0	0	0	2	1	0	0	0	0	0	0	0	1	10
Rear_ud_height	0	0	1	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	6
Strength	0	0	0	0	2	1	1	2	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	8
Body_depth	0	0	0	0	3	0	2	2	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	10
Stature	1	1	0	0	2	0	1	0	0	1	1	0	0	2	0	1	0	1	0	0	0	0	1	0	0	0	0	0	3	15
Rump_width	0	0	0	0	3	0	0	2	1	0	0	0	2	2	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	13
Foot_angle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2
Total	9	3	9	3	44	23	18	15	3	3	11	0	8	29	8	5	5	30	7	14	13	1	2	2	2	14	1	6	20	308

Table 4.4. A short list of candidate genes with high posterior probability of causality

GeneID	**GeneName**	**GeneChrom**	**GeneStart**	**GeneEnd**	**GeneType**	**Associated Traits**
536203	*ABCG2*	6	37902882	38030585	protein_coding	Fat|Fat_Percent|Milk|Net_Merit|Pro_Percent|Protein
100337258	*TMTC2*	5	12290927	12714498	protein_coding	Final_score|Fore_udder_att|Front_teat_pla|Rear_teat_pla|Rear_ud_height|Teat_length
541123	*ARRDC3*	7	93240415	93253099	protein_coding	Dtr_Calv_Ease|Rear_ud_height|Sire_Calv_Ease|Strength|Teat_length|Udder_depth
100336551	*ABCC9*	5	88672047	88834491	protein_coding	Dairy_form|Dtr_Preg_Rate|Fore_udder_att|Milk|Protein|Udder_depth
282609	*DGAT1*	14	1795425	1804838	protein_coding	Milk|Net_Merit|Pro_Percent|Protein|SCS
512656	*VPS13B*	14	66648395	67461111	protein_coding	Fat_Percent|Milk|Pro_Percent|Rear_ud_height|Udder_cleft
100125304	*ZNF613*	18	58100688	58141930	protein_coding	Body_depth|Net_Merit|Sire_Still_Birth|Stature|Strength
615414	*CCND2*	5	106253891	106276819	protein_coding	Body_depth|Rump_width|Stature|Strength
493719	*MGST1*	5	93925155	93950175	protein_coding	Fat|Fat_Percent|Milk|Pro_Percent
540317	*FGF6*	5	106157909	106169922	protein_coding	Body_depth|Rump_width|Stature|Strength
515039|107131630	*CCDC88C*	21	56629746	56773438	protein_coding	DFB_PTA|Dairy_form|Rear_ud_height
751788	*LOC751788*	5	102537290	102598580	other	Dairy_form|Final_score
280924	*SCD*	26	21137945	21148317	protein_coding	Fat|Fat_Percent

509011	*MKL1*	5	112260976	112472463	protein_coding	Milk|Protein
613562	*SYT8*	29	50287761	50294802	protein_coding	Final_score|Foot_angle
782261	*LOC782261*	14	1321274	1322712	protein_coding	Milk|Net_Merit
518897	*CHEK2*	17	70266805	70305258	protein_coding	Dtr_Calv_Ease|Sire_Calv_Ease
531757	*C8H9orf3*	8	82589563	83012157	protein_coding	Final_score|Rump_width
530076	*GC*	6	88687845	88739292	protein_coding	Cow_Conc_Rate|Udder_depth
540675	*KALRN*	1	69105208	69724961	protein_coding	Cow_Conc_Rate|Dtr_Preg_Rate
282208	*CSN1S1*	6	87141491	87159097	protein_coding	Pro_Percent|Protein
100140107	*SCAPER*	21	32118844	32548944	protein_coding	Fore_udder_att|Front_teat_pla
523297	*TCP11*	23	9018566	9067628	protein_coding	Stature|Udder_depth
280838	*PAEP*	11	103301488	103306381	protein_coding	Fat_Percent|Protein
527335	*ANKFN1*	19	7215828	7522300	protein_coding	Rump_width|SCS
513400	*NADSYN1*	29	48955458	48983419	protein_coding	Dtr_Preg_Rate|Stature
100852273	*LOC100852273*	15	49734992	49735929	protein_coding	Final_score|Fore_udder_att
616537	*RAB6A*	15	53936922	54027857	protein_coding	Milk|Pro_Percent
107132925	*LOC107132925*	11	38406991	38624018	lncRNA	Fore_udder_att|Udder_depth
281990	*POLD1*	18	57008175	57056561	protein_coding	Foot_angle|Protein
540709	*RAB11FIP2*	26	38617113	38657253	protein_coding	Front_teat_pla|Rear_teat_pla
616091	*MGMT*	26	49167460	49443160	protein_coding	Rump_angle
100141209	*BOSTAUV1R417*	18	58464593	58530789	protein_coding	Sire_Still_Birth
520463	*SLC50A1*	3	15518076	15520528	protein_coding	Pro_Percent
541287	*RNF217*	9	26436907	26577497	protein_coding	Pro_Percent
104974054	*LOC104974054*	14	39891940	40114343	lncRNA	Rump_angle
789567	*HSD17B12*	15	74652043	74830690	protein_coding	Fat_Percent

104975270	*LOC104975270*	20	33738712	33756847	lncRNA	Fore_udder_att
104972568	*LOC104972568*	5	107244108	107259980	lncRNA	Sire_Calv_Ease
537034	*ADGRV1*	7	92481179	92844786	protein_coding	Sire_Calv_Ease
508656	*CD276*	10	20323629	20355565	protein_coding	Dtr_Preg_Rate
537659	*TTC28*	17	69652292	70246650	protein_coding	Dtr_Calv_Ease
508832	*LSP1*	29	50238210	50277092	protein_coding	Udder_depth
100337421	*VEPH1*	1	110936431	111213545	protein_coding	Udder_cleft
615392	*TIGAR*	5	106223071	106238040	protein_coding	Prod_Life
518878	*CCDC57*	19	51271243	51381692	protein_coding	Fat
526125	*GON4L*	3	15004847	15093527	protein_coding	Protein
281152	*FASN*	19	51384892	51403614	protein_coding	Fat_Percent
504741	*COLEC12*	24	35630928	35816269	protein_coding	Rump_angle
507749	*C6*	20	33320064	33405582	protein_coding	SCS
317655	*MYH10*	19	28679649	28801223	protein_coding	Udder_depth
511614	*GPAT4*	27	36198042	36229006	protein_coding	Fat_Percent
616280	*EXOC6B*	11	11617983	12340418	protein_coding	Teat_length
515340	*ABO*	11	104231517	104270224	protein_coding	Pro_Percent
619012	*LOC619012*	29	39388027	39397193	pseudogene	Sire_Still_Birth
618771	*MRGPRG*	29	48989735	49027304	protein_coding	Sire_Calv_Ease
534482	*FSTL1*	1	65742626	65802423	protein_coding	Stature
282072	*SFTPD*	28	35814587	35824601	protein_coding	Pro_Percent
525618	*SLC24A2*	8	24495771	24782333	protein_coding	Rump_angle
407238	*ESR1*	9	89989608	90256185	protein_coding	Dtr_Calv_Ease
281276	*LDLR*	7	16768592	16802349	protein_coding	SCS

618784	*TBC1D22A*	5	118086396	118343834	protein_coding	Pro_Percent
520994	*PTCH1*	8	83518735	83581931	protein_coding	Body_depth
101903327	*LOC101903327*	14	7965390	8040409	lncRNA	Prod_Life
532711	*FAM98B*	10	34130152	34195283	protein_coding	Stature
530237	*VWA2*	26	34998522	35049697	protein_coding	Teat_length
786966	*LOC786966*	14	2054723	2089358	protein_coding	Pro_Percent
100140934	*MROH9*	16	39319532	39421012	protein_coding	Rear_teat_pla

Table 4.5. A short list of missense variants with posterior probability of causality of >0.1

Variant	Ref	Alt	Annotation	Gene	MAF	Average_PPC	Associated Traits
7:93244933	T	C	missense_variant	*ARRDC3*	0.10	0.608	Body_depth|Dtr_Calv_Ease|Net_Merit|Prod_Life|Rear_ud_height|Sire_Calv_Ease|Strength|Teat_length|Udder_depth
6:38027010	A	C	missense_variant	*ABCG2*	0.02	0.87	Fat|Fat_Percent|Milk|Net_Merit|Pro_Percent|Protein
8:85149325	C	A	missense_variant	*LOC101906801*	0.11	0.134	Body_depth|Final_score|Rump_width|Strength
21:56809835	G	A	missense_variant	*PPP4R3A*	0.01	0.191	Dairy_form|Prod_Life|Rear_ud_height
8:83581466	G	T	missense_variant	*PTCH1*	0.03	0.678	Body_depth|Strength
26:21144708	G	A	missense_variant&splice_region_variant	*SCD*	0.25	0.571	Fat|Fat_Percent
1:69673871	C	T	missense_variant	*KALRN*	0.11	0.462	Cow_Conc_Rate|Dtr_Preg_Rate
19:7521843	G	A	missense_variant	*ANKFN1*	0.22	0.446	Rump_width|SCS
29:50290087	G	A	missense_variant	*SYT8*	0.39	0.438	Final_score|Foot_angle
29:50286107	G	A	missense_variant	*TNNI2*	0.20	0.436	Rump_width|Stature
29:50289940	A	G	missense_variant	*SYT8*	0.39	0.399	Final_score|Foot_angle
17:70276788	G	A	missense_variant	*CHEK2*	0.09	0.388	Dtr_Calv_Ease|Sire_Calv_Ease
18:57017616	G	A	missense_variant	*POLD1*	0.10	0.291	Foot_angle|Protein
8:83044210	A	T	missense_variant	*FANCC*	0.12	0.252	Rear_teat_pla|Udder_depth
14:1321450	A	T	missense_variant	*LOC782261*	0.21	0.206	Milk|Net_Merit
5:67644905	G	A	missense_variant	*STAB2*	0.04	0.184	Body_depth|Teat_length
5:67677946	G	A	missense_variant	*STAB2*	0.04	0.184	Body_depth|Teat_length
7:19876364	C	T	missense_variant	*SAFB*	0.30	0.156	Body_depth|Stature
14:1321721	G	A	missense_variant	*LOC782261*	0.21	0.155	Milk|Net_Merit
5:68052261	C	G	missense_variant	*HCFC2*	0.04	0.145	Body_depth|Teat_length
14:1321349	T	G	missense_variant	*LOC782261*	0.21	0.143	Milk|Net_Merit

18:57521276	G	A	missense_variant	*CTU1*	0.06	0.099	DFB_PTA|Heifer_Conc_Rate
14:2072259	C	T	missense_variant& splice_region_variant	*LOC786966*	0.09	0.919	Pro_Percent
18:44378414	G	A	missense_variant	*CHST8*	0.12	0.889	DFB_PTA
26:22874498	C	T	missense_variant		0.18	0.827	Pro_Percent
5:118244695	C	T	missense_variant	*TBC1D22A*	0.18	0.676	Pro_Percent
5:30259026	G	A	missense_variant	*NCKAP5L*	0.25	0.611	Teat_length
3:15464749	G	A	missense_variant	*GBA*	0.06	0.601	Milk
3:20189903	G	A	missense_variant	*ADAMTSL4*	0.08	0.571	Dairy_form
11:104232298	C	T	missense_variant& splice_region_variant	*ABO*	0.31	0.449	Pro_Percent
19:51319797	A	G	missense_variant	*CCDC57*	0.35	0.423	Fat
18:61020273	C	T	missense_variant	*ZNF331*	0.04	0.322	Dairy_form
19:51319759	T	C	missense_variant	*CCDC57*	0.35	0.304	Fat
8:85147150	T	C	missense_variant	*LOC101906801*	0.12	0.302	Strength
13:58716308	G	A	missense_variant	*C13H20orf85*	0.12	0.297	Fore_udder_att
11:104232319	A	T	missense_variant	*ABO*	0.31	0.223	Pro_Percent
14:66328304	C	T	missense_variant	*SPAG1*	0.12	0.222	SCS
28:35824058	A	T	missense_variant	*SFTPD*	0.35	0.186	Pro_Percent
29:48976568	T	C	missense_variant	*NADSYN1*	0.02	0.183	Stature
29:48978814	G	A	missense_variant	*NADSYN1*	0.02	0.183	Stature
6:87181542	T	G	missense_variant	*CSN2*	0.05	0.181	Pro_Percent
29:50289452	C	T	missense_variant	*TNNI2*	0.08	0.158	Stature
11:103304757	T	C	missense_variant	*PAEP*	0.48	0.149	Protein
25:26381789	G	A	missense_variant	*SGF29*	0.08	0.122	Milk
25:26544685	T	A	missense_variant	*TAOK2*	0.08	0.119	Milk
25:26458669	G	A	missense_variant	*TBX6*	0.08	0.114	Milk

7:33638886	C	T	missense_variant	*LOC521901*	0.34	0.114	Rear_teat_pla
11:103303475	G	A	missense_variant	*PAEP*	0.48	0.11	Protein
28:18186635	A	G	missense_variant	*ARID5B*	0.35	0.105	Dairy_form

Figures

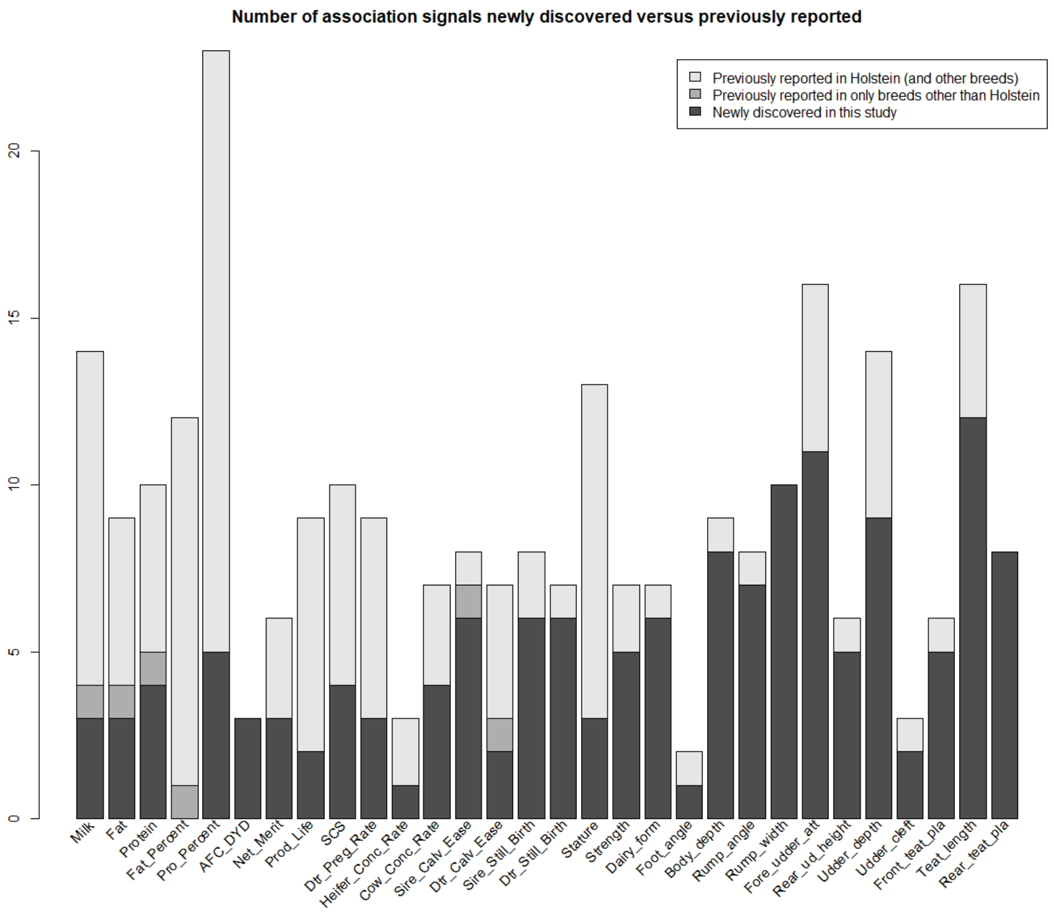


Figure 4.1. **Number of association signals newly discovered in our single-trait GWAS versus previously reported**. There are in total 30 traits listed. Three leg traits were not listed since we did not find associations passing whole-genome significance. Days to first breeding (DFB) and final score were not listed because there was no matched trait in the Cattle QTLdb release 35.

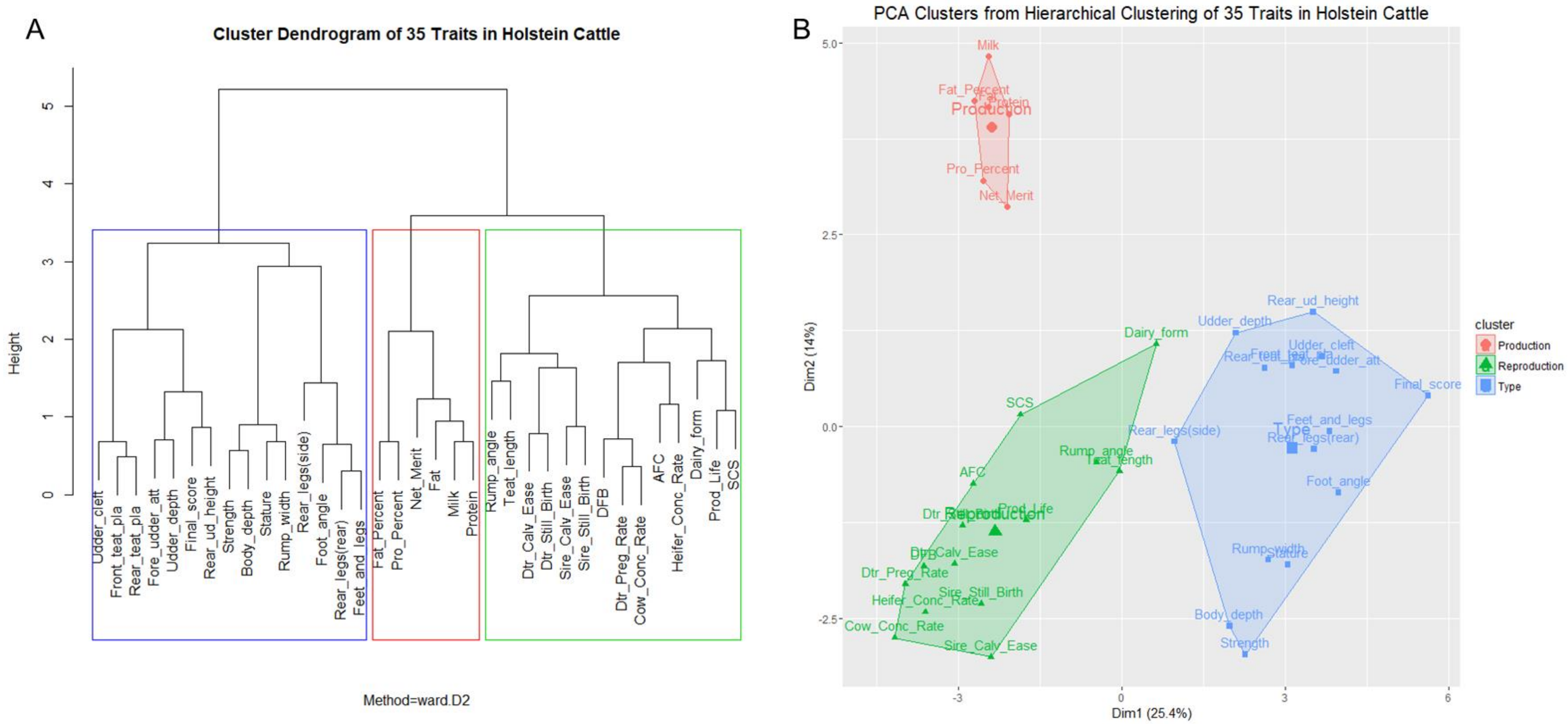


Figure 4.2. **Hierarchical clustering of 35 traits in Holstein cattle**. A. Cluster dendrogram. B. PCA clusters.

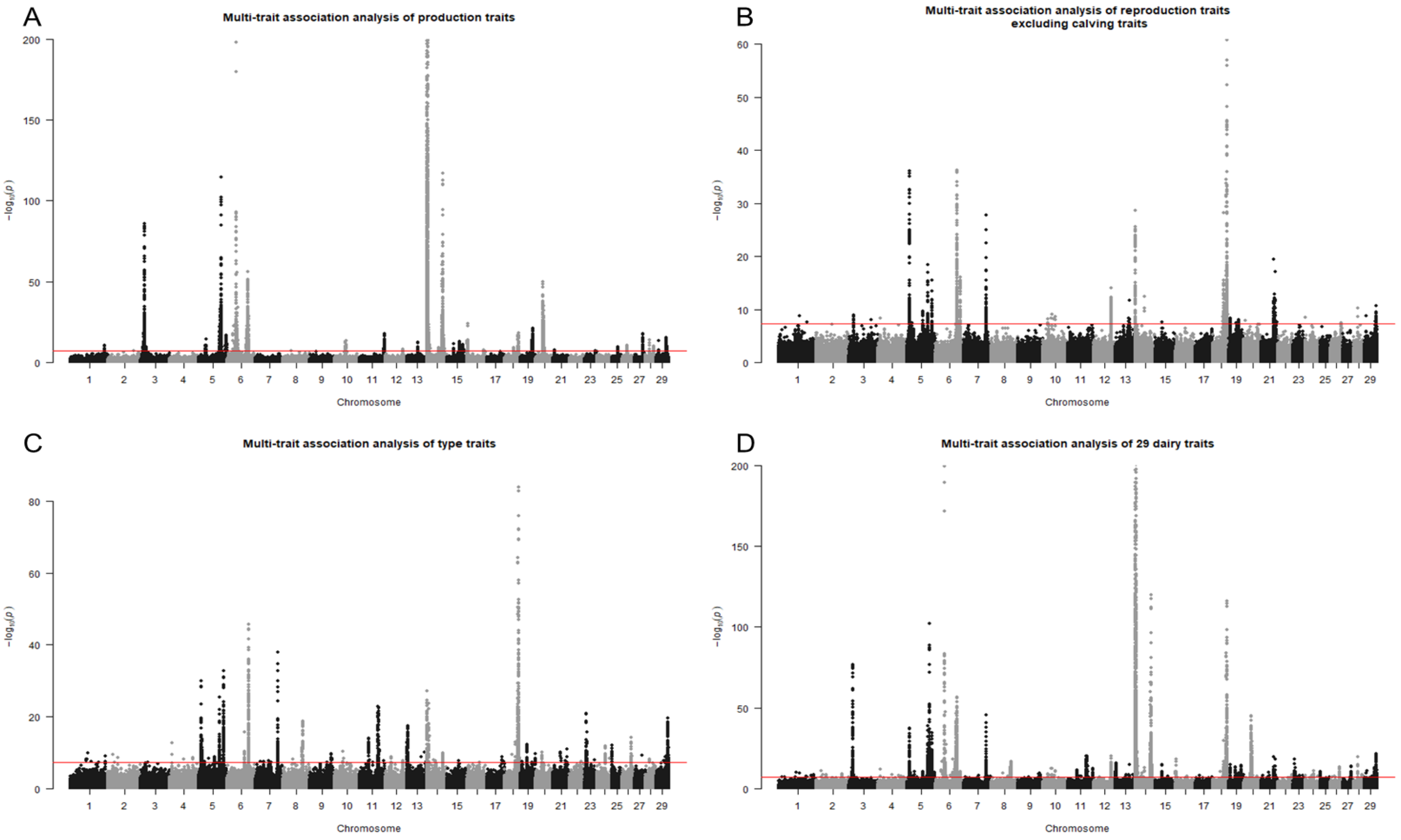


Figure 4.3. **Manhattan plots for multi-trait association analyses**. A. Production traits. B. Reproduction traits, excluding four calving traits (calving ease and stillbirth traits). C. Type traits. D. All 29 dairy traits, excluding days to first breeding (DFB), net merit and four calving traits. Red lines denote a significance of 5*E*-8.

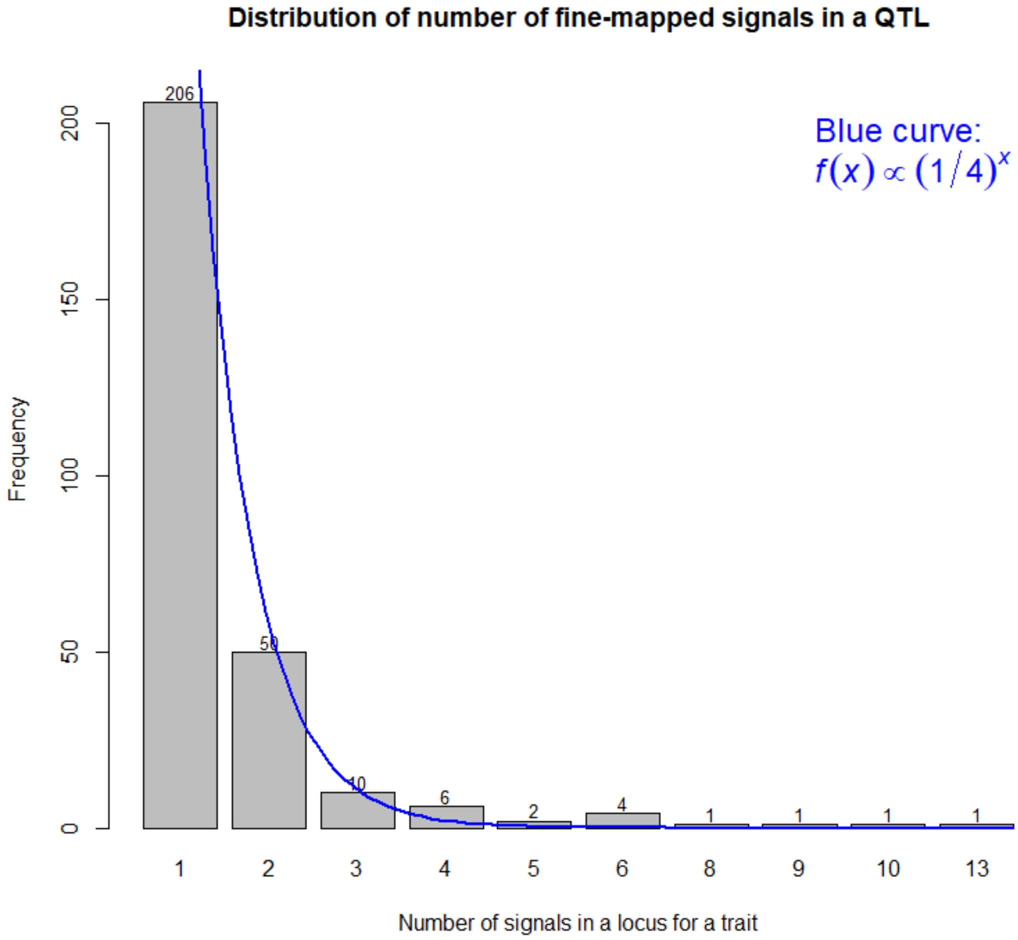


Figure 4.4. **Distribution of number of fine-mapped signals in a candidate locus for a trait**. Signals were filtered by a significance threshold of 5*E*-7.

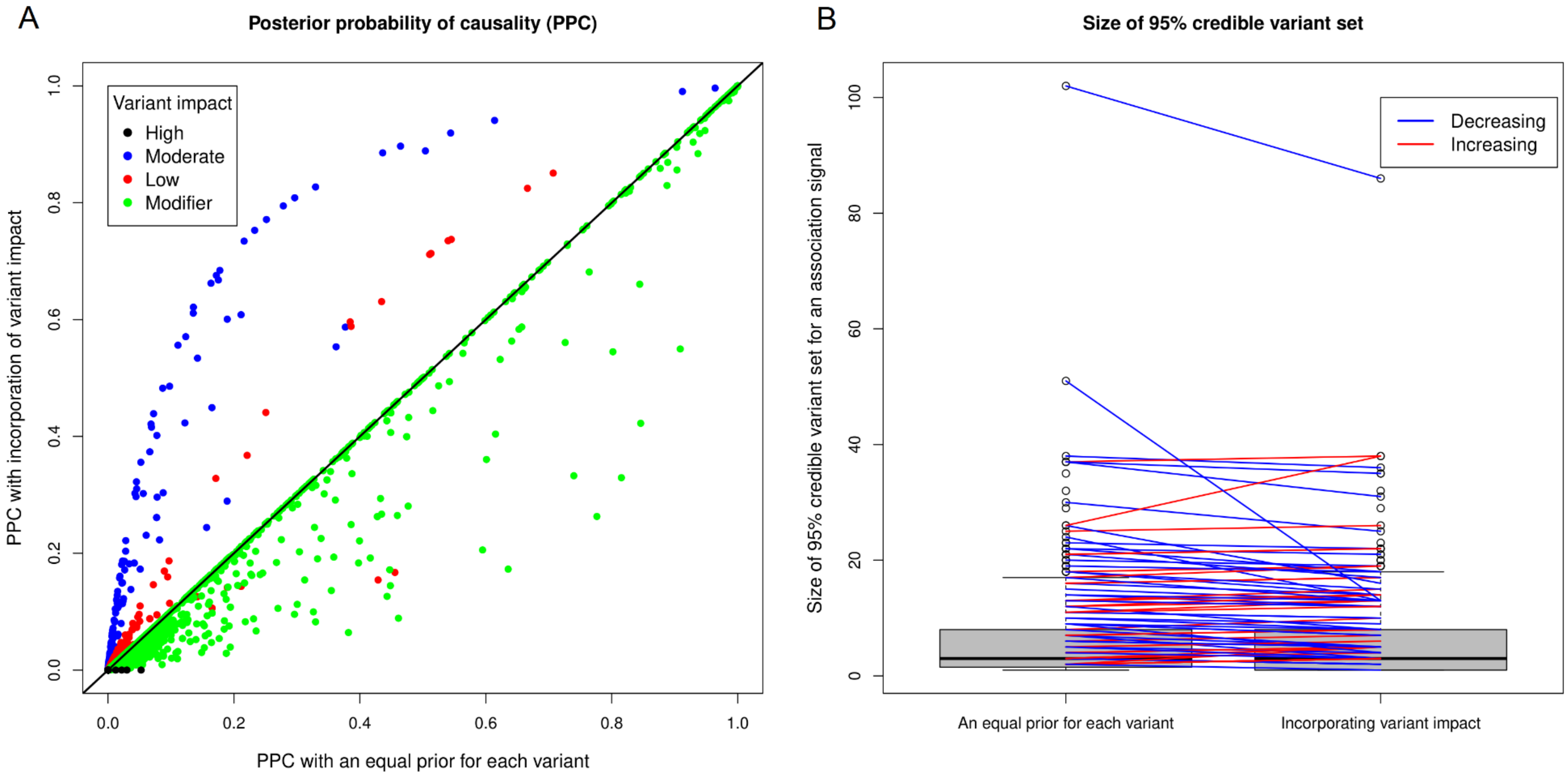


Figure 4.5. **Impact of incorporation of SnpEff-inferred effect impact on fine-mapping performance**. A. Posterior probability of causality (PPC) with incorporation of effect impact versus PPC with an equal prior for each variant. B. Size of 95% credible variant set overall decreased by incorporation of SnpEff-inferred effect impact.

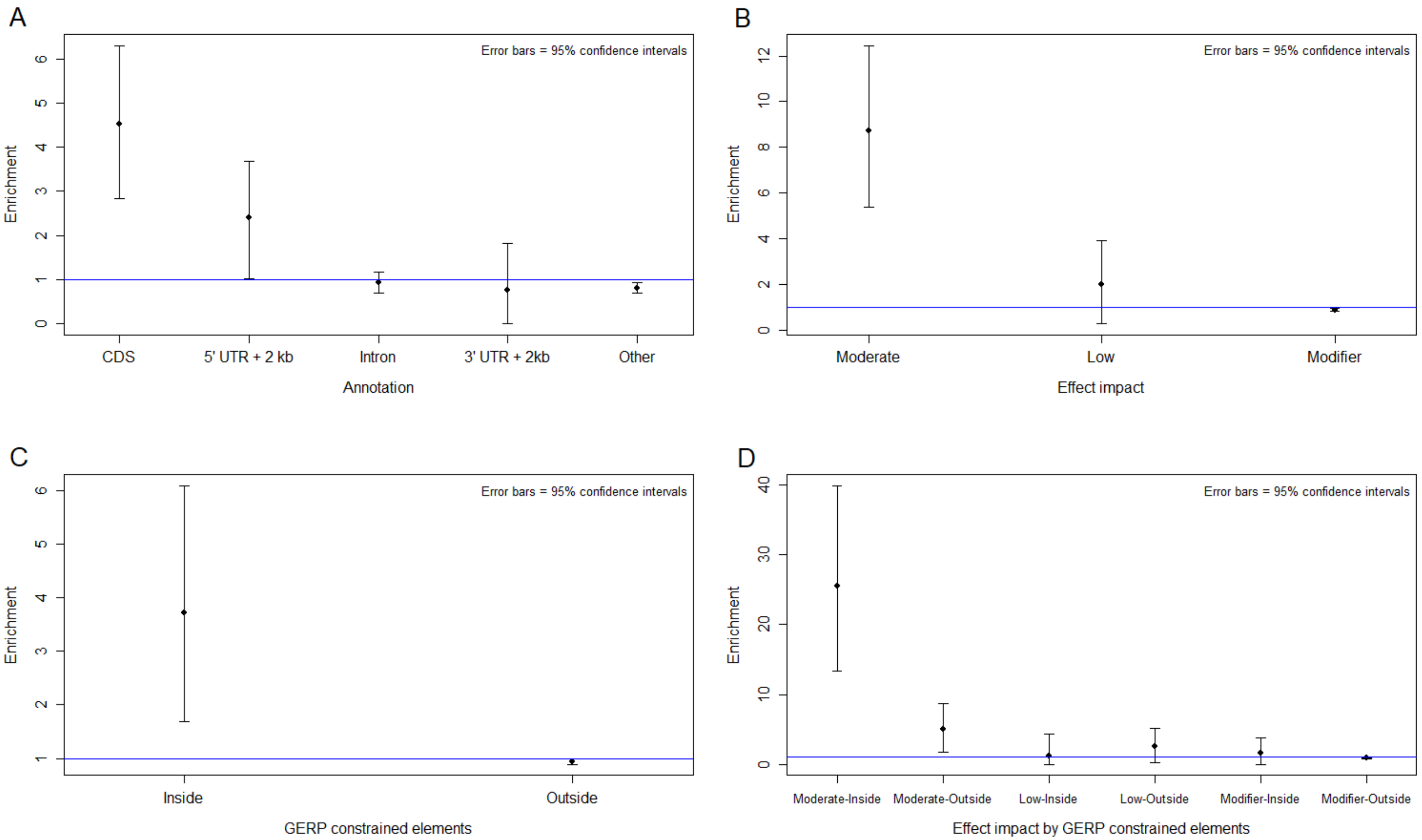


Figure 4.6. **Enrichment estimates for various functional annotations.** A. Locations of variants regarding protein-coding genes. B. SnpEff effect impact. C. GERP constrained elements. D. Effect impact by GERP constrained elements.

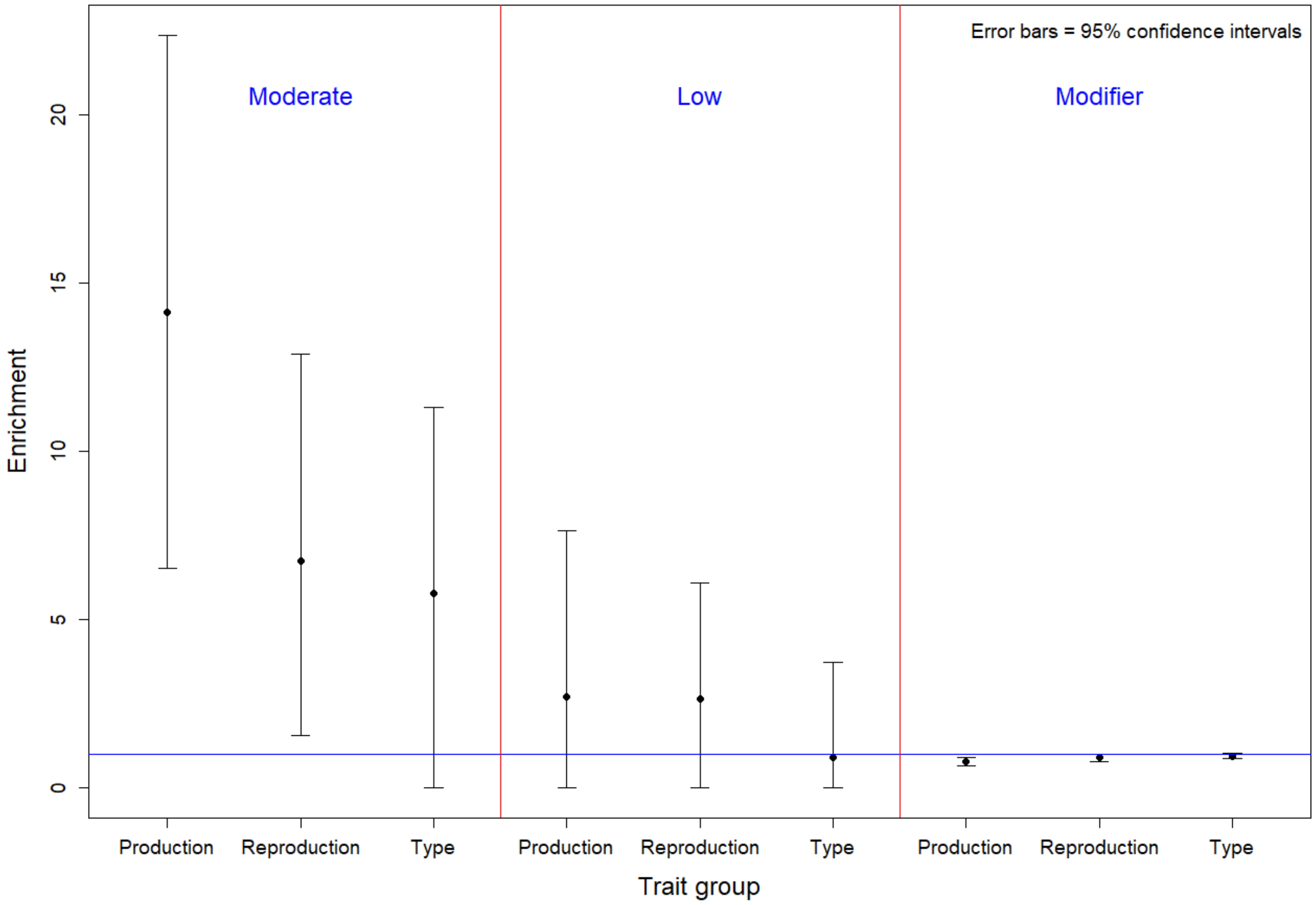


Figure 4.7. **Enrichment estimates for SnpEff effect impact obtained with three groups of traits separately**

Chapter 5: SNP-set based Genomic Prediction to Incorporate Functional Annotation

Abstract

Genomic prediction has emerged as an effective approach in plant and animal breeding. Including functional annotation into the genomic model can be of great advantage. Due to the statistical and computational challenges in large genomics studies, however, a fast and flexible method to incorporate such external information is still lacking. Here, we propose a Bayesian model that can incorporate functional annotation in a flexible way, implement two complementary algorithms to fit the model (namely, parameter expanded variational Bayes and Gibbs sampling), and develop a fast software package named SSGP. In our model, whole genome markers can be split into groups in a user-defined manner, and each group of markers is given a common effect variance. Since previous functional genomics studies have accumulated much evidence on which genes, genomic regions or pathways are more/less important for a trait of interest, we can divide genome-wide SNPs into a number of groups based on their levels of importance and then use the predefined SNP sets in SSGP. Additionally, each marker has a pre-specified weight for which the rule can be flexibly assigned, e.g. based on minor allele frequency or LD pattern. For testing purpose, we analyzed many data sets. Generally, SSGP could achieve similar prediction performance compared to the best approaches reported, though only proximity was used for grouping SNPs (markers were divided into continuous, non-overlapping chunks). It is also fast and capable of handling large data. Collectively, the method and software show great potential to increase accuracy in genomic prediction, particularly in the future when more useful functional annotations are becoming available.

Key words: genomic prediction, SNP set, functional annotation

Introduction

Genomic prediction (GP) has emerged as an effective approach in plant and animal breeding (Garcia-Ruiz *et al*, 2016). Most existing methods are solely based on mining marker genotypes and phenotypes (Habier *et al*, 2011; VanRaden, 2008), disregarding relevant information on biological mechanisms linking mutations to traits. Including functional annotation into the genomic model can be of great advantage. A straightforward way to achieve this is to group or weight SNP markers. Actually, such a way has been extensively used for partitioning heritability with multi-component GREML (Loh *et al*, 2015a; Yang *et al*, 2015) and for improving GRM (Speed *et al*, 2017). However, this method is computationally intractable when the number of groups is large. Additionally, it cannot directly generate SNP effect estimates, making it harder to predict phenotypes for new individuals.

To tackle the drawbacks of GREML for grouping and weighting markers, we here propose a Bayesian method, which can group variants in a manner similar to multiple-component GREML and weight variants in a user-defined manner. We implement two complementary algorithms to fit the model (namely, parameter expanded variational Bayes and Gibbs sampling), and develop a fast software package named SSGP. Extensive data analyses show that SSGP is fast and produces accurate predictions.

Methods

Statistical model

We use the following statistical model:

$$\begin{gathered}
\mathbf{y} = \mathbf{X}\mathbf{b} + \sum_{h=1}^{p} \mathbf{K}_h \mathbf{u}_h + \mathbf{e} \\
\mathbf{b} \sim N(\mathbf{0}, \mathbf{I}\sigma_b^2) \\
\mathbf{u}_h \sim N(\mathbf{0}, \mathbf{W}_h \sigma_{u_h}^2),\ h = 1, \cdots, p \\
\mathbf{e} \sim N(\mathbf{0}, \mathbf{R}\sigma_e^2) \\
\sigma_e^2 \sim \text{Inv-Gamma}(c_0, d_0) \\
\text{Prior 1: } \sigma_{u_h}^2 \sim \text{Inv-Gamma}(a_h, b_h),\ h = 1, \cdots, p \\
\text{Prior 2: } \sigma_{u_h} \sim \text{Half-Cauchy}(A_h),\ h = 1, \cdots, p
\end{gathered} \tag{5.1}$$

where the phenotype (***y***) is decomposed to three parts, namely the fixed effects (***b***), the random effects (***u**h*, *h=1,…,p*), and the residual (***e***). The fixed effects are assumed to follow a normal distribution with an extremely large variance that is pre-specified (e.g., $\sigma_b^2 = 1E8$). The residuals are assumed to follow a normal distribution, each having a pre-specified weight (i.e., element in the diagonal matrix **R**) for variance. There are *p* groups of random effects. The random effects within each group are assumed to follow normal distribution, and each group has its own variance component ($\sigma_{u_h}^2$) and pre-specified variant-specific weights (diagonal matrix $\mathbf{\Omega}_h$). Each variance component is further assumed to follow an inverse-gamma distribution with group-specific parameters (shape parameter a_h and scale parameter b_h) or a half-Cauchy distribution with scale A_h.

In the context of genomic prediction using SNP markers, the whole-genome markers are split into *p* groups, each group having a common variance of SNP effects

(Fig 5.1). In addition, one can specify weights for sequence variants based on previous functional enrichment analysis, like the result from S-LDSC (Finucane *et al*, 2015). Despite the difference in hyper-priors, model (5.1) is basically equivalent to the model of multiple-component GREML.

Algorithms for model fitting

We implement both variatonal Bayes (VB) (Beal, 2003) and Gibbs sampling to fit model (5.1). The two algorithms are complementary to each other. Variatonal Bayes is fast but produces irredeemably biased estimates, while Gibbs sampling is relatively slow but produces asymptotically unbiased estimates.

To speed up the convergence of VB, we applied the parameter expanded method to our VB iterations (Jaakkola and Qi, 2007). We expanded model (5.1) by introducing auxiliary variables (c_h, $h=1,\dots,p$) to each group of SNPs:

$$\hat{\mathbf{u}}_h = \mathbf{u}_h / c_h, \text{ and } \hat{\sigma}^2_{u_h} = \sigma^2_{u_h} / c^2_h . \tag{5.2}$$

In each iteration, after all parameters are updated, the variational lower bound is maximized with respect to c_h ($h=1,\dots,p$). $\mathbf{u}_h$ and $\sigma^2_{u_h}$ are then updated again using

$$\mathbf{u}_h = c_h \hat{\mathbf{u}}_h, \text{ and } \sigma^2_{u_h} = c^2_h \hat{\sigma}^2_{u_h} . \tag{5.3}$$

Gibbs sampler for the model with the half-Cauchy prior is also based on the use of auxiliary variables (Makalic and Schmidt, 2015).

The time complexity of one VB iteration is $O(s^2m + nm + nc)$, where n is sample size, s is the group size (if all groups are of equal size), m is the total number of all SNPs, and c is the number of covariates. Gibbs sampling is faster than VB for one iteration, but requires much more iterations than VB. In practice, VB is orders of magnitude faster than Gibbs sampling.

Software implementation

We develop the software tool in C++ with the Eigen 3 library for fast matrix computation and Intel MKL for fast random number generation. Our software tool is named SSGP (SNP-set based Genomic Prediction).

Example usage

Two examples on how to use SSGP are illustrated in Fig. 5.2. First, we can group SNPs of similar importance, as previous functional genomics studies have accumulated much evidence on which genes are more/less important for a trait of interest (Fig. 5.2A). A simple grouping way is to group SNPs based on proximity, in that SNPs close to each other tend to behave similarly due to LD. Second, we can weight SNPs based on their MAFs and LD scores (Speed *et al*, 2017), e.g., setting bigger weight to low-MAF SNPs (Fig. 5.2B).

Data analysis

Simulation data

We analyzed the 16th QTL-MAS workshop data. In this simulation data set, there are 3000 animals as training and 1200 as validation. Each animal has genotypes of ten thousand equally distributed SNPs. Fifty QTLs and three traits (milk, fat and % fat) are simulated.

Real data

We analyzed two traits (%CD4+ and %CD8+) in the WTCCC heterogeneous stock mice data set, which consists of ~1400 individuals and ~10k SNP markers (Valdar *et al*, 2006). %CD4+ and %CD8+ have a heritability of ~0.4 and ~0.9, respectively. We randomly split the sample into two equal parts, and used one part as training and the other as validation. The splitting was repeated 20 times.

We also analyzed five milk production traits in a large dairy cattle data set (VanRaden *et al*, 2017). The data set has genotypes of 760K SNPs. We used 20K old bulls as training and 4K young bulls as validation.

The human lipid profile data (Investigators, 1989) consist of ~10k unrelated individuals. We used 10-fold cross validation and ~620K whole-genome SNP markers to predict four lipid profile traits and body mass index.

Benchmarking

We compared SSGP to GBLUP (via GCTA) (Yang *et al*, 2011a), BayesA (VanRaden, 2008), BayesB (Nadaf *et al*, 2012), and BayesRv2 (Moser *et al*, 2015), in terms of prediction accuracy (or root-mean-square error (RMSE)), running speed and memory usage. External functional annotation was not used in SSGP. Instead, we just divided SNP markers into continuous, non-overlapping chunks (i.e., simple grouping based on proximity).

Results

Prediction accuracy or RMSE

QTL-MAS 2012 simulation data

When setting SNP-set size to 10 or 100, both MCMC and VB in SSGP performs better than BayesB which has been reported to be the best method for the data set (http://qtl-mas-2012.kassiopeagroup.com) (Fig. 5.3). In addition, MCMC produces higher prediction accuracy than VB, especially when SNP-set size is 1.

WTCCC heterogeneous stock mice data

BayesR and SSGP clearly produce smaller RMSEs than GBLUP and thus have better performance (Fig. 5.4). SSGP is overall similar to BayesR, but in some scenarios is significantly better (Fig. 5.4).

Dairy cattle data

In SSGP, each SNP set contains 1K continuous SNPs. SSGP-VB has an increase of up to 8 percentage points in prediction accuracy for the five milk production traits compared to BayesA (Fig. 5.5).

Human lipid profile data

For this data, each SNP set contains 200 continuous SNPs in SSGP. SSGP-VB has much higher prediction accuracy than GBLUP for all four lipid profile traits (Fig. 5.6). For body weight index, which is known to have a highly polygenic architecture, SSGP and GBLUP show similar genomic predictions (Fig. 5.6).

Speed and memory usage

SSGP-VB is faster than GBLUP by GCTA, even though the data sets have much smaller sample size than the number of SNPs and thus favor GCTA (Table 5.1). In addition, SSGP-VB is two orders of magnitude faster than BayesRv2, while SSGP-MCMC is slightly slower (Table 5.1). The speed of SSGP-MCMC is reasonable, considering that BayesRv2 uses an improved algorithm for updating effects across multiple SNPs in blocks (Calus, 2014) and is one of the fastest MCMC-based computing tools for genomic prediction. As shown in Table 5.2, SSGP is also memory efficient.

Both time cost and memory usage in SSGP are linearly proportional to sample size and number of markers. It is accordingly projected that SSGP can complete genomic prediction of one trait for two million animals and 60K SNPs in one day with a few cores of modern computer processor.

Conclusion

We propose a flexible method to incorporate functional annotation into genomic prediction, and develop a fast software tool, SSGP. SSGP can readily handle very large data sets. The method and software show great potential to increase accuracy in genomic prediction. Our data analyses also show that SNP grouping based on proximity is helpful.

References

Beal MJ (2003). *Variational algorithms for approximate Bayesian inference*.

Calus MP (2014). Right-hand-side updating for fast computing of genomic breeding values. *Genet Sel Evol* **46:** 24.

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR *et al* (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**(11)**:** 1228-1235.

Garcia-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-Lopez FJ, Van Tassell CP (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc Natl Acad Sci U S A* **113**(28)**:** E3995-4004.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011). Extension of the Bayesian alphabet for genomic selection. *BMC bioinformatics* **12**(1)**:** 186.

Investigators A (1989). The atherosclerosis risk in communit (aric) study: Design and objectwes. *American journal of epidemiology* **129**(4)**:** 687-702.

Jaakkola TS, Qi Y. (2007). *Advances in Neural Information Processing Systems*, pp 1097-1104.

Loh PR, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ *et al* (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet* **47**(12)**:** 1385-1392.

Makalic E, Schmidt DF (2015). A simple sampler for the horseshoe estimator. *arXiv preprint arXiv:150803884*.

Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM (2015). Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *Plos Genetics* **11**(4).

Nadaf J, Riggio V, Yu TP, Pong-Wong R (2012). Effect of the prior distribution of SNP effects on the estimation of total breeding value. *BMC Proc* **6 Suppl 2:** S6.

Speed D, Cai N, Consortium U, Johnson MR, Nejentsev S, Balding DJ (2017). Reevaluation of SNP heritability in complex human traits. *Nat Genet* **49**(7)**:** 986-992.

Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO *et al* (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature genetics* **38**(8)**:** 879-887.

VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**(11)**:** 4414-4423.

VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet Sel Evol* **49**(1)**:** 32.

Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH *et al* (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47**(10)**:** 1114-1120.

Yang J, Lee SH, Goddard ME, Visscher PM (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**(1)**:** 76-82.

Tables

Table 5.1. Running speed of GCTA, BayesRv2 and SSGP

Trait	Data size		Time cost (minutes)			
	Samples	SNPs	GCTA[b]	BayesRv2[cd]	SSGP-VB	SSGP-MCMC[d]
TG[a]	8,240	612,926	18.6 (0.1)	NA	12.3 (0.2)	NA
%CD4+	704	9,159	0.0417 (0.0085)	4.15 (0.23)	0.0283 (0.0095)	7.39 (0.067)

[a] GCTA and SSGP were used with 10 cores of Intel Xeon E5-2680 v2.

[b] The time used by GCTA included time for building GRM, GREML and calculating SNP effects.

[c] BayesRv2 was used with options –msize 500 and –blocksize 2.

[d] The chain length was 50,000.

Table 5.2. Peak memory usage (Gb)

Trait	Data size		Peak memory usage (Gb)	
	Samples	SNPs	GCTA-GREML	SSGP
TG	8,240	612,926	1.7	~6

TG: Triglycerides.

Figures

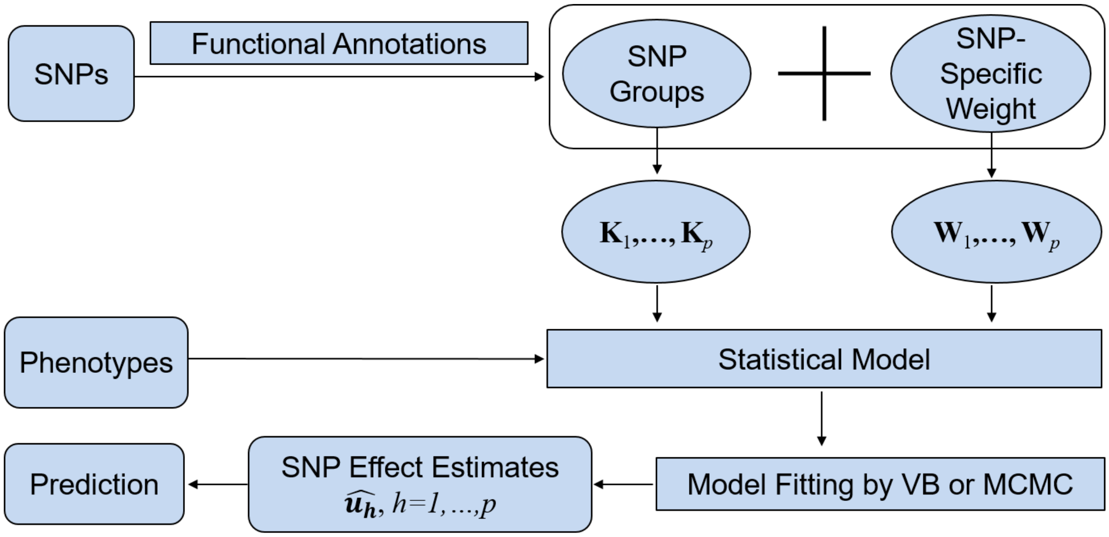
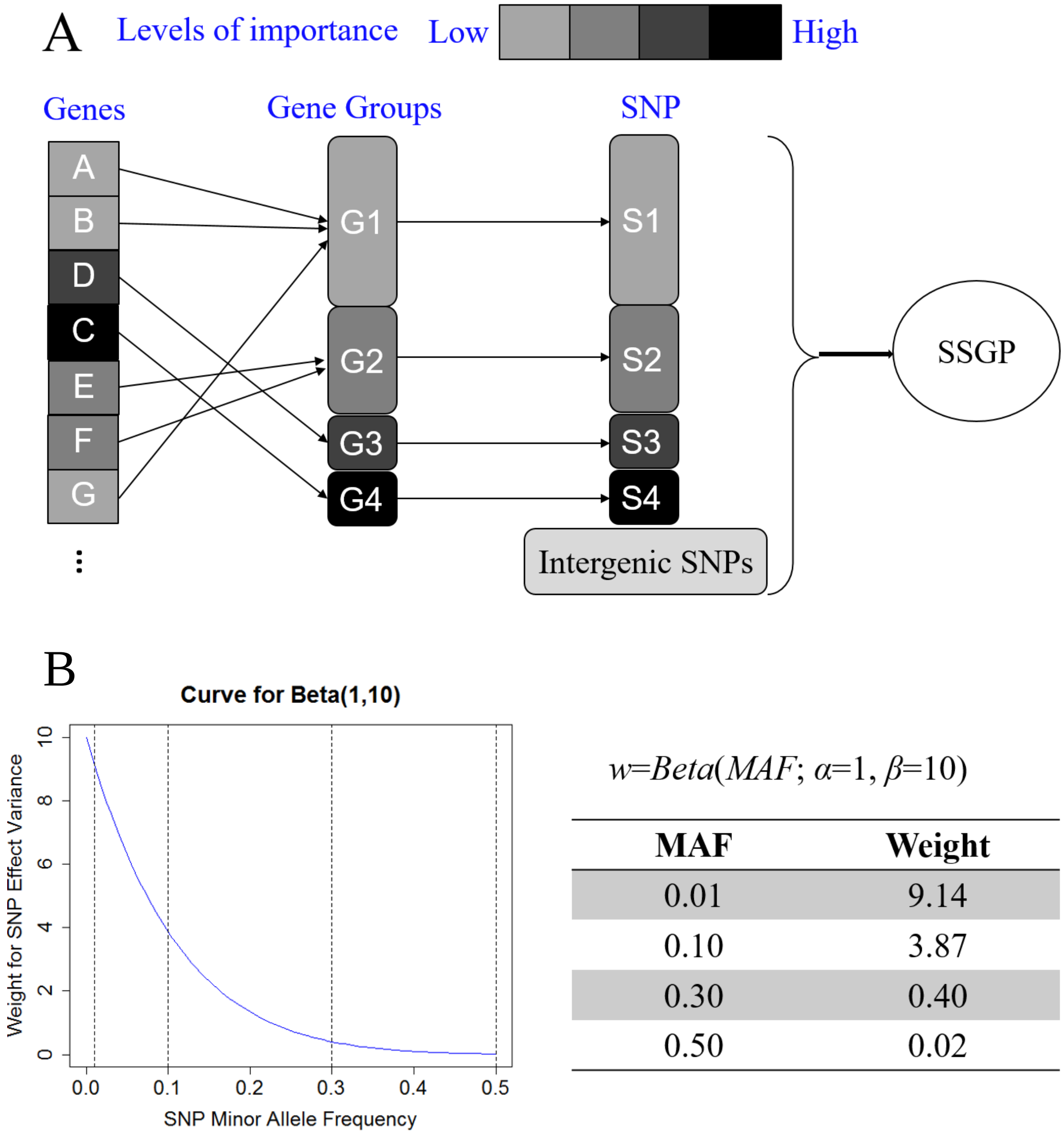

SNPs
Functional Annotations
SNP Groups
SNP-Specific Weight
$\mathbf{K}_1,\ldots,\mathbf{K}_p$
$\mathbf{W}_1,\ldots,\mathbf{W}_p$
Phenotypes
Statistical Model
Prediction
SNP Effect Estimates
$\widehat{\boldsymbol{u}_{\boldsymbol{h}}}, h=1,\ldots,p$
Model Fitting by VB or MCMC

Figure 5.1. Scheme of incorporating functional annotations into genomic prediction by SSGP. Symbols in the figure are the same as in model (5.1).



$w = Beta(MAF; \alpha=1, \beta=10)$

MAF	Weight
0.01	9.14
0.10	3.87
0.30	0.40
0.50	0.02

Figure 5.2. Two examples on how to use SSGP.

A) Grouping SNPs based on relative importance

B) Weighting SNPs based on MAFs

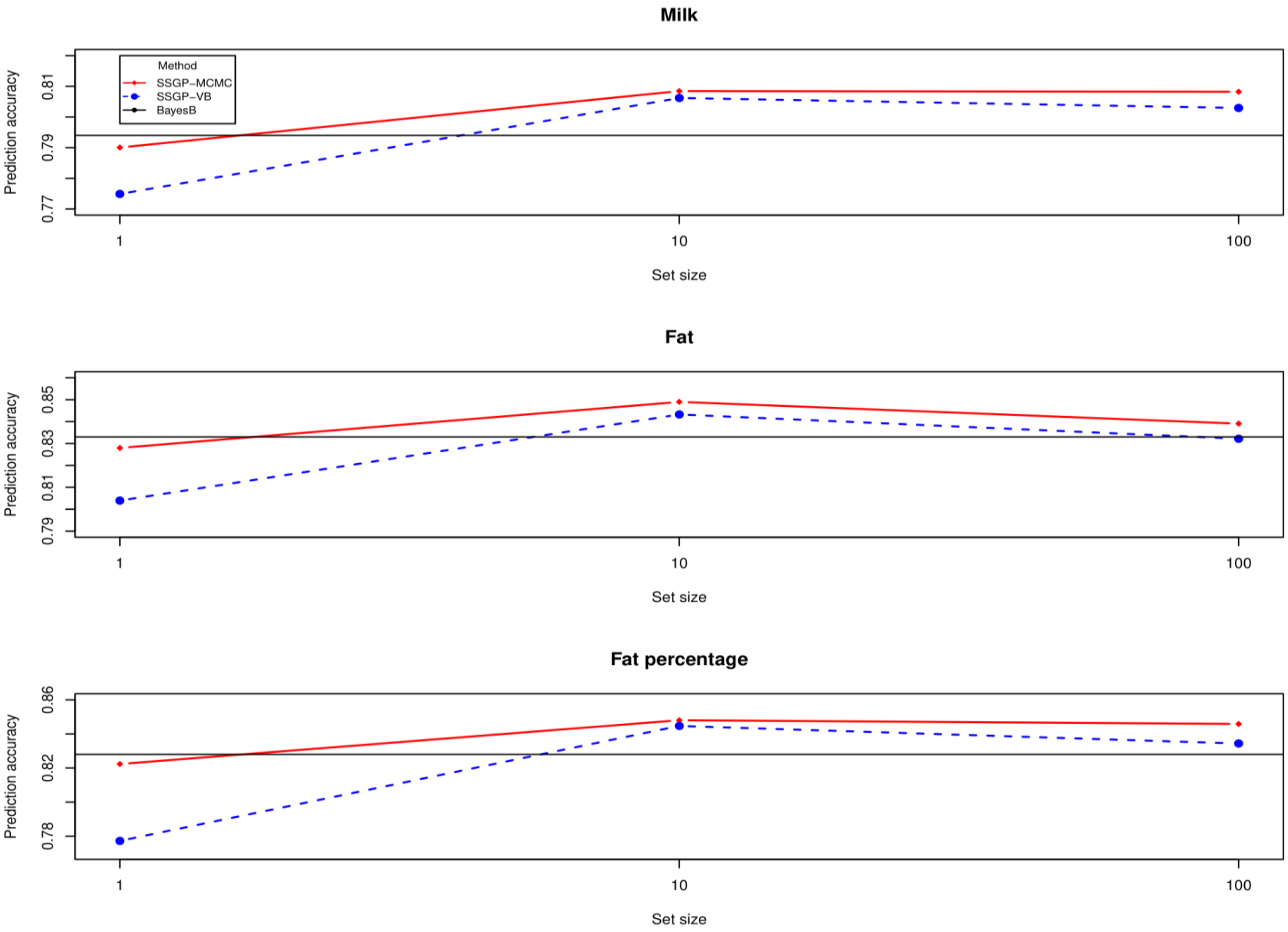


Figure 5.3. Prediction accuracies of SSGP and BayesB for the QTL-MAS 2012 simulation data

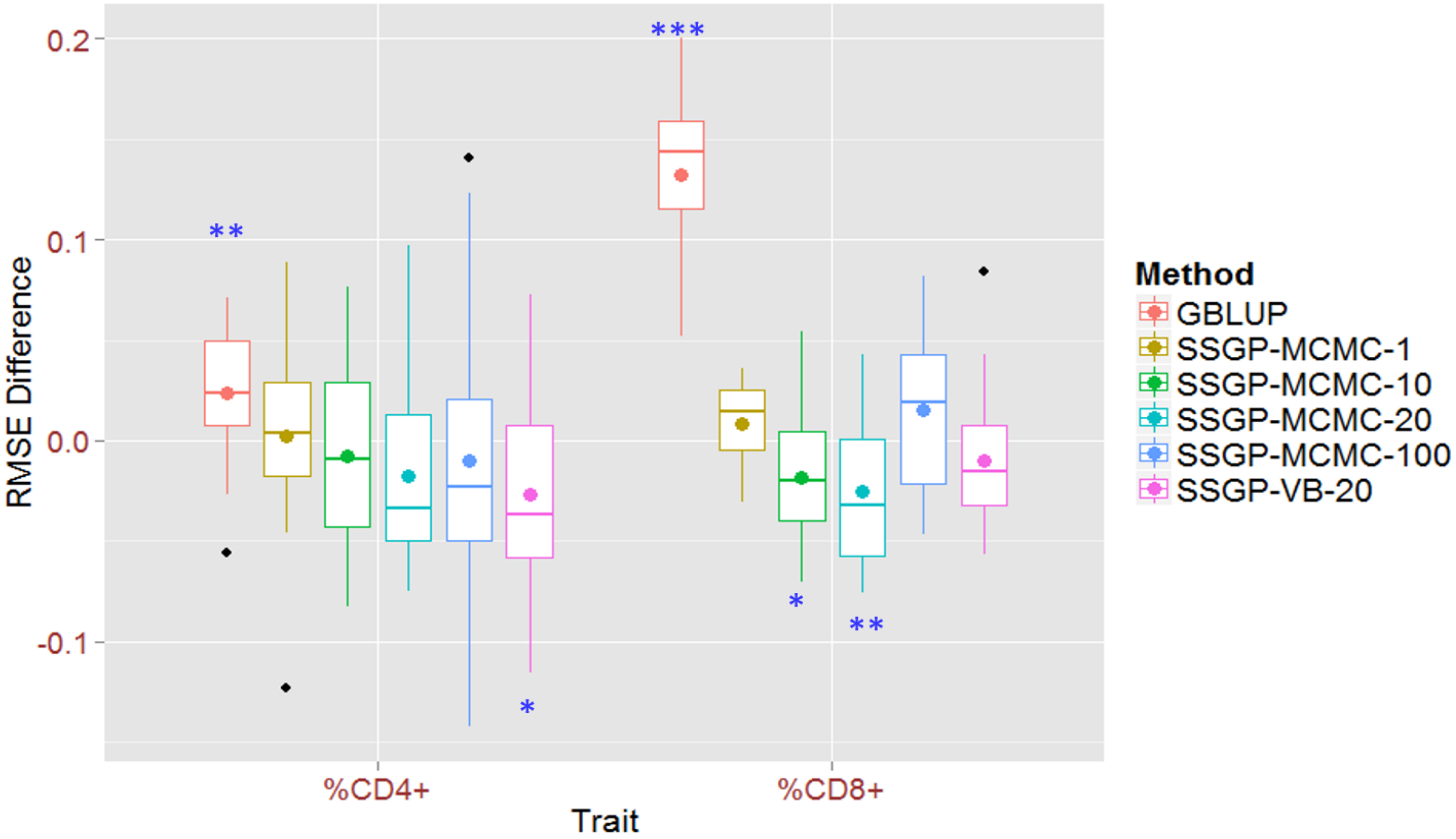


Figure 5.4. Root-mean-square errors (RMSEs) of GBLUP and SSGP compared to BayesR for the WTCCC heterogeneous stock mice data. The baseline is BayesR. Boxplots show the RMSE difference of GBLUP/SSGP as compared to BayesR. Negative values indicate a better performance than BayesR. SSGP-MCMC-1, -10, -20, and -100 denote a SNP set size of 1, 10, 20, and 100, respectively. *: *p*-value<0.05. **: *p*-value<0.01. ***: *p*-value<0.001. The *p*-values are from the tests for whether the RMSE difference equals 0.

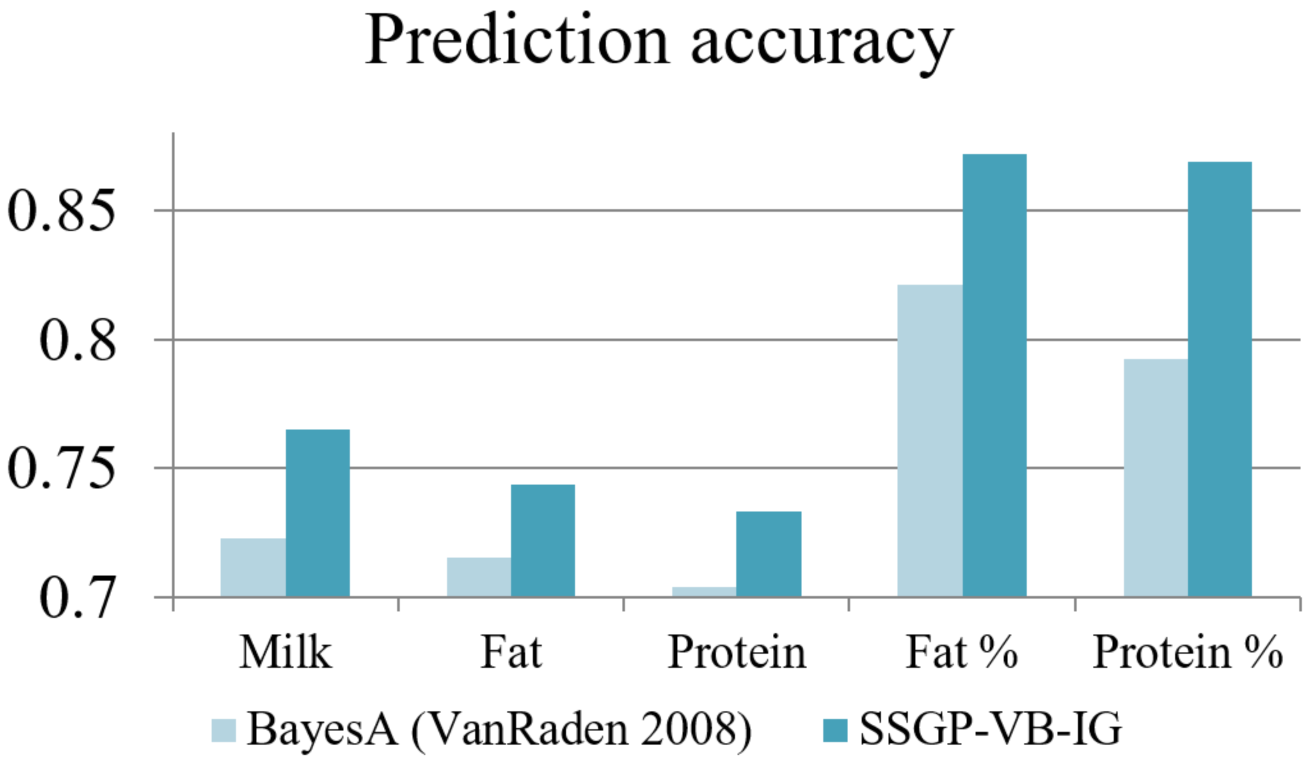


Figure 5.5. Prediction accuracies of BayesA and SSGP-VB for five milk production traits. SSGP-VB-IG denotes SSGP VB with an inverse-gamma prior.

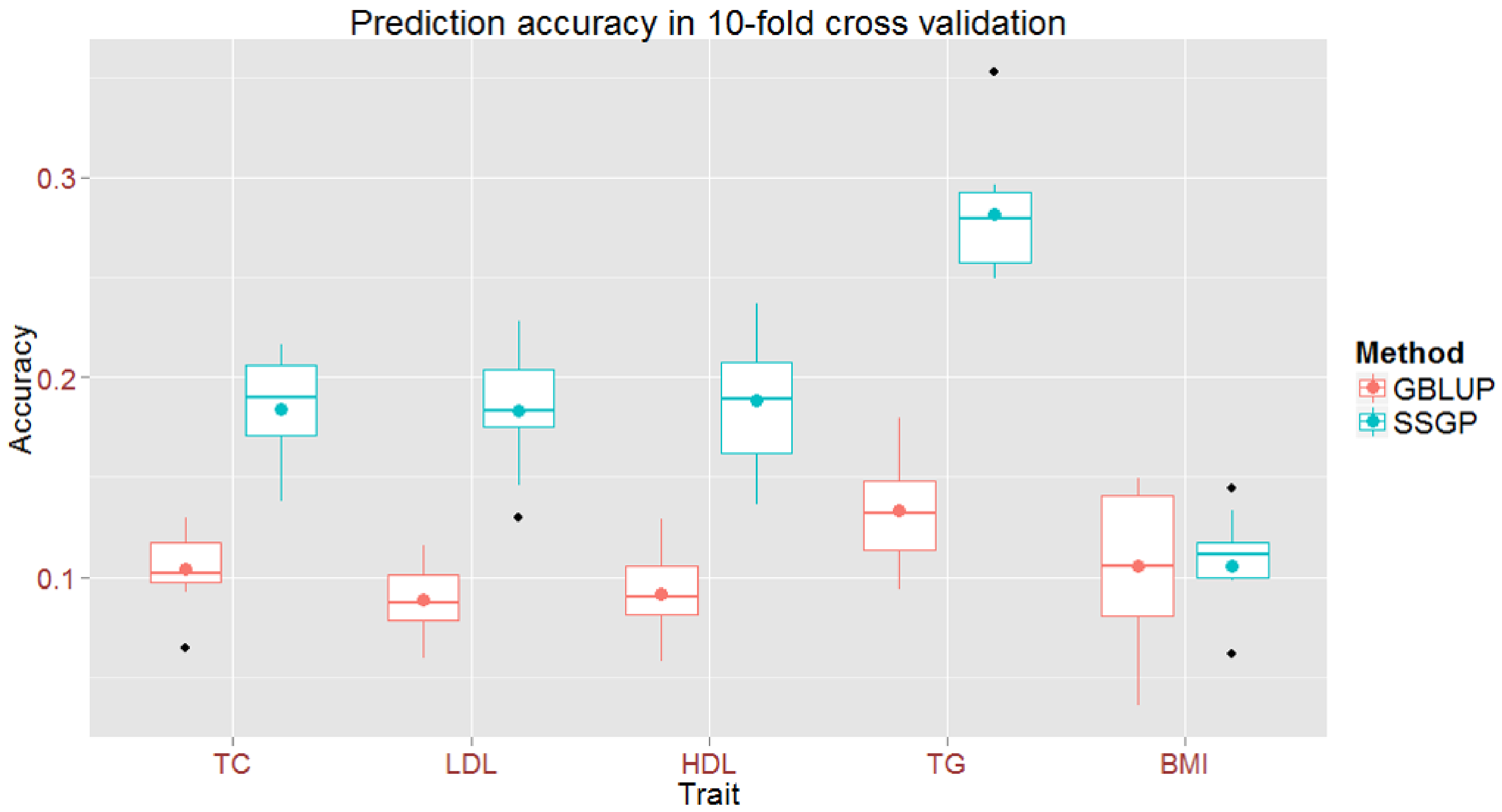


Figure 5.6. Prediction accuracies of GBLUP and SSGP for four lipid profile traits and body mass index. GBLUP was performed using GCTA. SSGP was run using VB. TC: Total cholesterol. LDL: Low-density lipoprotein. HDL: High-density lipoprotein. TG: Triglycerides. BMI: Body mass index.

Chapter 6: Conclusions

The objective of this research was to gain knowledge on the genetic architecture of complex traits and to develop a method for using the knowledge to improve genomic prediction in dairy cattle. Studies in Chapters 2-5 were all centered on this objective.

In **Chapter 2**, we aimed to dissect additive and non-additive genetic effects for production, reproduction and health traits in dairy cattle. By the study, we have found that non-additive effects contributed a non-negligible amount (more for reproduction traits) to the total genetic variance of complex traits in cattle. We also identified a dominance QTL for milk yield, demonstrating that detection of QTLs with non-additive effect is possible in GWAS using a large dataset.

In **Chapter** 3, I aimed to develop a powerful method and a fast software tool for SNP-set association and fine-mapping. In the study, I proposed a unified Bayesian model for single-marker/SNP-set association and fine-mapping and developed a software tool, BFMAP, which can deal with both population and pedigree data. I demonstrated that BFMAP achieves a power similar to or higher than existing software tools but is at least a few times faster with respect to single-marker/SNP-set association tests. I also showed that BFMAP performs well for fine-mapping even for complex linkage disequilibrium structures. Additionally, BFMAP can easily incorporate functional annotation into fine-mapping and efficiently use fine-mapping results to do functional enrichment analysis. Our method and software tool will be especially useful for unraveling causal effect enrichment patterns, as many more functional annotations are becoming available in dairy cattle genome.

In **Chapter 4**, we aimed to identify QTLs underlying the complex traits in Holstein cattle using imputed sequence data, and to fine-map 35 production, reproduction, and body conformation traits to single-gene resolution. By the study, we found many novel association signals and identified many promising candidate genes, including some previously reported ones. We also showed causal effect enrichment patterns for a few functional annotations available in dairy cattle genome and demonstrated that our fine-mapping result can be readily used for future functional studies. This study may facilitate follow-up functional validation and expand our understanding of complex traits in dairy cattle.

In **Chapter 5**, I aimed to develop an efficient method and a fast computing tool for using functional annotations in genomic prediction. In the study, I proposed a Bayesian model that can incorporate functional annotation in a flexible way, implemented both variational Bayes and Gibbs sampling to fit the model, and developed a fast software package named SSGP. I illustrated how to use SSGP to incorporate functional annotation in genomic prediction. I also demonstrated by extensive data analyses that the method and software have great potential to increase accuracy in genomic prediction and the capability to handle very large data.

It should be noted that the studies in these four chapters are closely related with each other and can be further integrated together. This directly provides a future direction. For example, the causal effect enrichment patterns in the Chapter 4 study can be readily used in SSGP to test a functional annotation-driven GP model for dairy cattle. The tests for non-additive effects in the Chapter 2 study can be readily improved by BFMAP.

It should also be noted that BFMAP and SSGP are applicable for any species. As sequence data are rapidly growing for many livestock species, fine-mapping to single-gene or even single-variant resolution is becoming feasible. BFMAP will be especially useful for these studies, in that it has features favorable to livestock data. In addition, as the FAANG or other related projects produce more functional annotations on animal genomes, BFMAP will be also useful for discovering causal effect enrichment patterns. Furthermore, the discovered enrichment patterns can be readily used in SSGP to test more sophisticated genomic prediction models driven by functional annotations.

In the near future, I am particularly interested in testing prediction accuracy of SSGP for current dairy cattle genomic evaluation data maintained at the CDCB. In the Chapter 5 study, SSGP showed a considerable increase in prediction accuracy compared to BayesA (the method currently used in CDCB evaluations) when sequence genotypes were used. In that analysis, we did not use any functional annotation. Instead, we grouped markers based on only their proximity. It is interesting to see whether the proximity-based marker grouping also benefits 60K SNP genotypes which are currently used in practice. If the advantage is still available, it will be possible to apply SSGP to dairy cattle breeding considering its capability of handling very big data.

Appendix A

Proof of the Equivalence between Scaling Genotypes and Weighting Variants

Suppose that the weight of variance for variant *i* is A_{ii}. Here we prove that weighting variants via **A** is equivalent to scaling genotypes by square root of corresponding weights. For unscaled genotypes **Z**, we compute the scaled genotypes (denoted by $\tilde{\mathbf{Z}}$) by $\tilde{\mathbf{Z}} = \mathbf{Z}\mathbf{A}^{1/2}$. Based on equation (3.2), it is easy to obtain $\log P\left(\mathbf{Z}, \mathbf{A}, \mathbf{X}, \mathbf{y} \middle| M\right) = \log P\left(\tilde{\mathbf{Z}}, \mathbf{X}, \mathbf{y} \middle| M\right)$. Thus, weighting variants results in the same Bayes factor for any marker set as scaling genotypes by square root of corresponding weights.

Next, we show that they also result in the same null distribution of Bayes factor. For **Z**, we have $\mathbf{H} = \mathbf{P}\mathbf{Z}\left(\mathbf{Z}'\mathbf{P}'\mathbf{P}\mathbf{Z} + \gamma^{-1}\mathbf{A}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{P}'$ where **H**, **P**, and γ are the same as in equation (3.9). Similarly, we have $\tilde{\mathbf{H}} = \mathbf{P}\tilde{\mathbf{Z}}\left(\tilde{\mathbf{Z}}'\mathbf{P}'\mathbf{P}\tilde{\mathbf{Z}} + \gamma^{-1}\right)^{-1}\tilde{\mathbf{Z}}'\mathbf{P}'$ for $\tilde{\mathbf{Z}}$. Therefore, $\mathbf{H} = \tilde{\mathbf{H}}$, which results in the same null distribution according to equation (3.9).

Bibliography

Abo-Ismail MK, Brito LF, Miller SP, Sargolzaei M, Grossi DA, Moore SS *et al* (2017). Genome-wide association studies and genomic prediction of breeding values for calving performance and body conformation traits in Holstein cattle. *Genet Sel Evol* **49**(1)**:** 82.

Alavi MV, Bette S, Schimpf S, Schuettauf F, Schraermeyer U, Wehrl HF *et al* (2007). A splice site mutation in the murine Opa1 gene features pathology of autosomal dominant optic atrophy. *Brain* **130**(Pt 4)**:** 1029-1042.

Aliloo H, Pryce J, González-Recio O, Cocks B, Goddard M, Hayes B (2016). Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. *Journal of Dairy Science*.

Aliloo H, Pryce JE, González-Recio O, Cocks BG, Hayes BJ (2016). Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. *Genetics Selection Evolution* **48**(1)**:** 8.

Álvarez-Castro JM (2015). Dissecting genetic effects with imprinting. *Models and Estimation of Genetic Effects***:** 35.

Ambros V (2004). The functions of animal microRNAs. *Nature* **431**(7006)**:** 350-355.

Amin N, van Duijn CM, Aulchenko YS (2007). A genomic background based method for association analysis in related individuals. *PLoS One* **2**(12)**:** e1274.

Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW *et al* (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol* **16:** 57.

Aulchenko YS, de Koning DJ, Haley C (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**(1)**:** 577-585.

Bartlett MS (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44**(3-4)**:** 533-534.

Beal MJ (2003). *Variational algorithms for approximate Bayesian inference*.

Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**(10)**:** 1493-1501.

Bohmanova J, Sargolzaei M, Schenkel FS (2010). Characteristics of linkage disequilibrium in North American Holsteins. *BMC Genomics* **11:** 421.

Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K *et al* (2014). A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS Genet* **10**(3)**:** e1004198.

Brotherstone S, Goddard M (2005). Artificial selection and maintenance of genetic variance in the global dairy cow population. *Philos Trans R Soc Lond B Biol Sci* **360**(1459)**:** 1479-1488.

Browning BL, Browning SR (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**(2)**:** 210-223.

Browning BL, Browning SR (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**(2)**:** 459-471.

Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR *et al* (2015). An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**(11)**:** 1236-1241.

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C *et al* (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**(3)**:** 291-295.

Caballero A, Tenesa A, Keightley PD (2015). The nature of genetic variation for complex traits revealed by GWAS and regional heritability mapping analyses. *Genetics***:** genetics. 115.177220.

Calus MP (2014). Right-hand-side updating for fast computing of genomic breeding values. *Genet Sel Evol* **46:** 24.

Calus MP, Veerkamp RF (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution* **43**(1)**:** 26.

Carlborg Ö, Haley CS (2004). Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics* **5**(8)**:** 618-625.

Cebamanos L, Gray A, Stewart I, Tenesa A (2014). Regional heritability advanced complex trait analysis for GPU and traditional parallel architectures. *Bioinformatics* **30**(8)**:** 1177-1179.

Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA *et al* (2015). Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**(3)**:** 719-736.

Chen W, McDonnell SK, Thibodeau SN, Tillmans LS, Schaid DJ (2016). Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics. *Genetics* **204**(3)**:** 933-958.

Cheng H, Kizilkaya K, Zeng J, Garrick D, Fernando R (2018). Genomic Prediction from Multiple-Trait Bayesian Regression Methods Using Mixture Priors. *Genetics* **209**(1)**:** 89-103.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L *et al* (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**(2)**:** 80-92.

Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee JH *et al* (2005). Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res* **15**(7)**:** 936-944.

Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S *et al* (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**(7)**:** 901-913.

Cunningham F, Moore B, Ruiz-Schultz N, Ritchie GR, Eilbeck K (2015). Improving the Sequence Ontology terminology for genomic variant annotation. *J Biomed Semantics* **6:** 32.

Da Y, Wang C, Wang S, Hu G (2014). Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS One* **9**(1)**:** e87666.

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF *et al* (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46**(8)**:** 858-865.

Davies RB (1980). Algorithm AS 155: The Distribution of a Linear Combination of Chi-Square Random Variables. *Journal of the Royal Statistical Society Series C (Applied Statistics)* **29**(3)**:** 323-333.

De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H *et al* (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **312**(5777)**:** 1215-1217.

Dean A (2011). In the loop: long range chromatin interactions and gene regulation. *Brief Funct Genomics* **10**(1)**:** 3-10.

Decker JE (2015). Agricultural Genomics: Commercial Applications Bring Increased

Basic Research Power. *PLoS Genet* **11**(11)**:** e1005621.

Devlin B, Roeder K (1999). Genomic control for association studies. *Biometrics* **55**(4)**:** 997-1004.

Eu-Ahsunthornwattana J, Howey RA, Cordell HJ (2014). Accounting for relatedness in family-based association studies: application to Genetic Analysis Workshop 18 data. *BMC Proc* **8**(Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo)**:** S79.

Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S *et al* (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**(7539)**:** 337.

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR *et al* (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**(11)**:** 1228-1235.

Foster DP, George EI (1994). The Risk Inflation Criterion for Multiple-Regression. *Ann Stat* **22**(4)**:** 1947-1975.

Gao N, Martini JWR, Zhang Z, Yuan X, Zhang H, Simianer H *et al* (2017). Incorporating Gene Annotation into Genomic Prediction of Complex Phenotypes. *Genetics* **207**(2)**:** 489-501.

Garcia-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-Lopez FJ, Van Tassell CP (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc Natl Acad Sci U S A* **113**(28)**:** E3995-4004.

Garrick DJ, Taylor JF, Fernando RL (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol* **41:** 55.

Gianola D (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics***:** genetics. 113.151753.

Gianola D, Fernando RL, Stella A (2006). Genomic assisted prediction of genetic value with semi-parametric procedures. *Genetics*.

Grisart B, Farnir F, Karim L, Cambisano N, Kim JJ, Kvasz A *et al* (2004). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc Natl Acad Sci U S A* **101**(8)**:** 2398-2403.

Guo X, Christensen OF, Ostersen T, Wang Y, Lund MS, Su G (2016). Genomic prediction using models with dominance and imprinting effects for backfat thickness and average daily gain in Danish Duroc pigs. *Genetics Selection Evolution* **48**(1)**:** 67.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011). Extension of the Bayesian alphabet for genomic selection. *BMC bioinformatics* **12**(1)**:** 186.

Hans C, Dobra A, West M (2007). Shotgun Stochastic search for "Large p" regression. *Journal of the American Statistical Association* **102**(478)**:** 507-516.

Hao X, Zeng P, Zhang S, Zhou X (2018). Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS Genet* **14**(1)**:** e1007186.

Henderson CR (1984). *Applications of linear models in animal breeding*. Guelph : University of Guelph.

Hendricks AE, Dupuis J, Logue MW, Myers RH, Lunetta KL (2014). Correction for multiple testing in a gene region. *Eur J Hum Genet* **22**(3)**:** 414-418.

Hill WG, Goddard ME, Visscher PM (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* **4**(2)**:** e1000008.

Hofer A (1998). Variance component estimation in animal breeding: a review. *Journal of Animal Breeding and Genetics* **115**(1-6)**:** 247-265.

Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* **198**(2)**:** 497-508.

Hou L, Zhao H (2013). A review of post-GWAS prioritization approaches. *Front Genet* **4:** 280.

Hu Y, Rosa GJ, Gianola D (2015). A GWAS assessment of the contribution of genomic imprinting to the variation of body mass index in mice. *BMC genomics* **16**(1)**:** 576.

Hu ZL, Park CA, Reecy JM (2016). Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Res* **44**(D1)**:** D827-833.

Huang H, Fang M, Jostins L, Umicevic Mirkov M, Boucher G, Anderson CA *et al* (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**(7662)**:** 173-178.

Investigators A (1989). The atherosclerosis risk in communit (aric) study: Design and objectwes. *American journal of epidemiology* **129**(4)**:** 687-702.

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X (2013). Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet* **92**(6)**:** 841-853.

Jaakkola TS, Qi Y. (2007). *Advances in Neural Information Processing Systems*, pp 1097-1104.

Jia Y, Jannink J-L (2012). Multiple trait genomic selection methods increase genetic value prediction accuracy. *Genetics***:** genetics. 112.144246.

Jiang J, Shen B, O'Connell JR, VanRaden PM, Cole JB, Ma L (2017). Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. *BMC Genomics* **18**(1)**:** 425.

Jiang J, Zhang Q, Ma L, Li J, Wang Z, Liu J (2015). Joint prediction of multiple quantitative traits using a Bayesian multivariate antedependence model. *Heredity* **115**(1)**:** 29.

Johnson D, Thompson R (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *Journal of dairy science* **78**(2)**:** 449-456.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB *et al* (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**(4)**:** 348-354.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ *et al* (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**(3)**:** 1709-1723.

Kemper KE, Goddard ME (2012). Understanding and predicting complex traits: knowledge from cattle. *Hum Mol Genet* **21**(R1)**:** R45-51.

Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL *et al* (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**(10)**:** e1004722.

Kim ES, Kirkpatrick BW (2009). Linkage disequilibrium in the North American Holstein population. *Anim Genet* **40**(3)**:** 279-288.

Kuonen D (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**(4)**:** 929-935.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP (2008). A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet* **82**(2)**:** 386-397.

Lee SH, Wray NR, Goddard ME, Visscher PM (2011). Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**(3)**:** 294-305.

Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**(19)**:** 2540-2542.

Legarra A, Christensen OF, Aguilar I, Misztal I (2014). Single Step, a general approach for genomic selection. *Livestock Science* **166:** 54-65.

Legarra A, Robert-Granié C, Croiseau P, Guillaume F, Fritz S (2011). Improved Lasso for genomic selection. *Genetics research* **93**(1)**:** 77-87.

Lewin B (2008). *Genes 9*. Jones & Bartlett Learning.

Li J, Ji L (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**(3)**:** 221-227.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011). FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**(10)**:** 833-835.

Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D (2012). Improved linear mixed models for genome-wide association studies. *Nat Methods* **9**(6)**:** 525-526.

Liu Z, Jaitner J, Reinhardt F, Pasman E, Rensing S, Reents R (2008). Genetic evaluation of fertility traits of dairy cattle using a multiple-trait animal model. *J Dairy Sci* **91**(11)**:** 4333-4343.

Loh PR, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ *et al* (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet* **47**(12)**:** 1385-1392.

Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM *et al* (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**(3)**:** 284-290.

Lopes MS, Bastiaansen JW, Janss L, Knol EF, Bovenhuis H (2015). Estimation of additive, dominance, and imprinting genetic variance using genomic data. *G3: Genes| Genomes| Genetics* **5**(12)**:** 2629-2637.

Lourenco DAL, Fragomeni BO, Bradford HL, Menezes IR, Ferraz JBS, Aguilar I *et al* (2017). Implications of SNP weighting on single-step genomic predictions for different reference population sizes. *J Anim Breed Genet* **134**(6)**:** 463-471.

Ma L, Brautbar A, Boerwinkle E, Sing CF, Clark AG, Keinan A (2012). Knowledge-Driven Analysis Identifies a Gene-Gene Interaction Affecting High-Density Lipoprotein Cholesterol Levels in Multi-Ethnic Populations. *PLoS genetics* **8**(5).

Ma L, O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C *et al* (2015). Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLoS genetics* **11**(11)**:** e1005387.

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E *et al* (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**(D1)**:** D896-D901.

Madsen BE, Browning SR (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**(2)**:** e1000384.

Makalic E, Schmidt DF (2015). A simple sampler for the horseshoe estimator. *arXiv preprint arXiv:150803884*.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ *et al* (2009). Finding the missing heritability of complex diseases. *Nature* **461**(7265)**:** 747-753.

Masuda Y, Misztal I, Tsuruta S, Legarra A, Aguilar I, Lourenco DAL *et al* (2016). Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. *J Dairy Sci* **99**(3)**:** 1968-1974.

Meister G, Tuschl T (2004). Mechanisms of gene silencing by double-stranded RNA. *Nature* **431**(7006)**:** 343-349.

Mendell JT, Dietz HC (2001). When the message goes awry: disease-producing mutations that influence mRNA content and performance. *Cell* **107**(4)**:** 411-414.

Meuwissen TH, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4)**:** 1819-1829.

Misztal I, Aggrey SE, Muir WM (2013). Experiences with a single-step genome evaluation. *Poult Sci* **92**(9)**:** 2530-2534.

Misztal I, Tsuruta S, Aguilar I, Legarra A, VanRaden PM, Lawlor TJ (2013). Methods to approximate reliabilities in single-step genomic evaluation. *J Dairy Sci* **96**(1)**:** 647-654.

Moore T, Haig D (1991). Genomic imprinting in mammalian development: a parental tug-of-war. *Trends in Genetics* **7**(2)**:** 45-49.

Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM (2015). Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *Plos Genetics* **11**(4).

Nadaf J, Riggio V, Yu TP, Pong-Wong R (2012). Effect of the prior distribution of SNP effects on the estimation of total breeding value. *BMC Proc* **6 Suppl 2:** S6.

Nayeri S, Sargolzaei M, Abo-Ismail MK, May N, Miller SP, Schenkel F *et al* (2016). Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. *BMC Genet* **17**(1)**:** 75.

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**(4)**:** e1000888.

Nishio M, Satoh M (2015). Genomic best linear unbiased prediction method including imprinting effects for genomic evaluation. *Genetics Selection Evolution* **47**(1)**:** 32.

O'Connell JR (2015). MMAP User Guide. Available: http://edn.som.umaryland.edu/mmap/index.php. Accessed 8 October 2015.

O'Connell JR. (2013). *63th Annual Meeting of The American Society of Human Genetics*.

O'Sullivan BP, Freedman SD (2009). Cystic fibrosis. *Lancet* **373**(9678)**:** 1891-1904.

Owens TW, Rogers RL, Best SA, Ledger A, Mooney AM, Ferguson A *et al* (2014). Runx2 is a novel regulator of mammary epithelial cell fate in development and breast cancer. *Cancer research* **74**(18)**:** 5277-5286.

Pal LR, Yu CH, Mount SM, Moult J (2015). Insights from GWAS: emerging landscape of mechanisms underlying complex trait disease. *BMC Genomics* **16 Suppl 8:** S4.

Patterson HD, Thompson R (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**(3)**:** 545-554.

Pegolo S, Cecchinato A, Mele M, Conte G, Schiavon S, Bittante G (2016). Effects of candidate gene polymorphisms on the detailed fatty acids profile determined by gas chromatography in bovine milk. *J Dairy Sci* **99**(6)**:** 4558-4573.

Saatchi M, Schnabel RD, Taylor JF, Garrick DJ (2014). Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. *BMC Genomics* **15:** 442.

Servin B, Stephens M (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**(7)**:** e114.

Shirali M, Pong-Wong R, Navarro P, Knott S, Hayward C, Vitart V *et al* (2016). Regional heritability mapping method helps explain missing heritability of blood lipid traits in isolated populations. *Heredity (Edinb)* **116**(3)**:** 333-338.

Spain SL, Barrett JC (2015). Strategies for fine-mapping complex traits. *Human molecular genetics* **24**(R1)**:** R111-R119.

Speed D, Cai N, Consortium U, Johnson MR, Nejentsev S, Balding DJ (2017). Reevaluation of SNP heritability in complex human traits. *Nat Genet* **49**(7)**:** 986-992.

Speed D, Hemani G, Johnson MR, Balding DJ (2012). Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* **91**(6)**:** 1011-1021.

Stranden I, Garrick DJ (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J Dairy Sci* **92**(6)**:** 2971-2975.

Su G, Christensen OF, Ostersen T, Henryon M, Lund MS (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS One* **7**(9)**:** e45293.

Sun C, VanRaden P, O'Connell J, Weigel K, Gianola D (2013). Mating programs including genomic relationships and dominance effects. *Journal of dairy science* **96**(12)**:** 8014-8023.

Sun C, VanRaden PM, Cole JB, O'Connell JR (2014). Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. *PLoS One* **9**(8)**:** e103934.

Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A, Masson G *et al* (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet* **48**(3)**:** 314-317.

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012). Rapid variance components-based method for whole-genome association analysis. *Nat Genet* **44**(10)**:** 1166-1170.

Team RC (2013). R: A language and environment for statistical computing.

Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO *et al* (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet* **38**(8)**:** 879-887.

van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I *et al* (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol* **46:** 41.

VanRaden P (2016). Practical implications for genetic modeling in the genomics era. *Journal of dairy science* **99**(3)**:** 2405-2412.

VanRaden PM (1986). Computational strategies for estimation of variance components. *Retrospective Theses and Dissertations*. 8319.

VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**(11)**:** 4414-4423.

VanRaden PM. (2016). *Vol. 2016*: Animal Improvement Program, Animal Genomics and Improvement Laboratory, ARS, USDA. .

VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB *et al* (2013). Genomic imputation and evaluation using high-density Holstein genotypes. *J Dairy Sci* **96**(1)**:** 668-678.

VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet Sel Evol* **49**(1)**:** 32.

VanRaden PM, Wiggans GR (1991). Derivation, calculation, and use of national animal model information. *J Dairy Sci* **74**(8)**:** 2737-2746.

Varona L, Legarra A, Toro MA, Vitezica ZG (2018). Non-additive Effects in Genomic Selection. *Front Genet* **9:** 78.

Vitezica ZG, Varona L, Legarra A (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* **195**(4)**:** 1223-1230.

Wakefield J (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genet Epidemiol* **33**(1)**:** 79-86.

Wang B, Sverdlov S, Thompson E (2017). Efficient Estimation of Realized Kinship from Single Nucleotide Polymorphism Genotypes. *Genetics* **205**(3)**:** 1063-1078.

Wang C, Prakapenka D, Wang S, Pulugurta S, Runesha HB, Da Y (2014). GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. *BMC bioinformatics* **15**(1)**:** 270.

Wittenburg D, Melzer N, Reinsch N (2015). Genomic additive and dominance variance of milk performance traits. *Journal of Animal Breeding and Genetics* **132**(1)**:** 3-8.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ *et al* (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* **86**(6)**:** 929-942.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**(1)**:** 82-93.

Xiang T, Christensen OF, Vitezica ZG, Legarra A (2016). Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. *Genetics Selection Evolution* **48**(1)**:** 92.

Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH *et al* (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47**(10)**:** 1114-1120.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR *et al* (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**(7)**:** 565-569.

Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, Replication DIG *et al* (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**(4)**:** 369-375, S361-363.

Yang J, Lee SH, Goddard ME, Visscher PM (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**(1)**:** 76-82.

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM *et al* (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* **43**(6)**:** 519-525.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**(2)**:** 100-106.

Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet* **49**(9)**:** 1304-1310.

Yang W, Tempelman RJ (2011). A Bayesian antedependence model for whole genome prediction. *Genetics***:** genetics. 111.131540.

Zhang L, Liu Y, Song F, Zheng H, Hu L, Lu H *et al* (2011). Functional SNP in the microRNA-367 binding site in the 3'UTR of the calcium channel ryanodine receptor gene 3 (RYR3) affects breast cancer risk and calcification. *Proc Natl Acad Sci U S A* **108**(33)**:** 13653-13658.

Zhang X, Cowper-Sal lari R, Bailey SD, Moore JH, Lupien M (2012). Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. *Genome Res* **22**(8)**:** 1437-1446.

Zhou Q, Guan Y (2017). On the Null Distribution of Bayes Factors in Linear Regression. *Journal of the American Statistical Association***:** 1-10.

Zhou X, Stephens M (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**(7)**:** 821-824.

Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N *et al* (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am J Hum Genet* **96**(1)**:** 21-36.

Zhu Z, Bakshi A, Vinkhuyzen AA, Hemani G, Lee SH, Nolte IM *et al* (2015). Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet* **96**(3)**:** 377-385.

Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE *et al* (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**(5)**:** 481-487.