J. Dairy Sci. 106:5218–5241
https://doi.org/10.3168/jds.2022-22694

Invited review: Good practices in genome-wide association studies to identify candidate sequence variants in dairy cattle

G. Sahana,[1]* Z. Cai,[1] M. P. Sanchez,[2] A. C. Bouwman,[3] and D. Boichard[2]
[1]Aarhus University, Center for Quantitative Genetic and Genomics, 8830 Tjele, Denmark
[2]Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France
[3]Wageningen University & Research, Animal Breeding and Genomics, 6700 AH Wageningen, the Netherlands

ABSTRACT

Genotype data from dairy cattle selection programs have greatly facilitated GWAS to identify variants related to economic traits. Results can enhance the accuracy of genomic prediction, analyze more complex models that go beyond additive effects, elucidate the genetic architecture of a trait, and finally, decipher the underlying biology of traits. The entire process, comprising data generation, quality control, statistical analyses, interpretation of association results, and linking results to biology should be designed and executed to minimize the generation of false-positive and false-negative associations and misleading links to biological processes. This review aims to provide general guidelines for data analysis that address data quality control, association tests, adjustment for population stratification, and significance evaluation to improve the reliability of conclusions. We also provide guidance on post-GWAS strategy and the interpretation of results. These guidelines are tailored to dairy cattle, which are characterized by long-range linkage disequilibrium, large half-sib families, and routinely collected phenotypes, requiring different approaches than those applied in human GWAS. We discuss common limitations and challenges that have been overlooked in the analysis and interpretation of GWAS to identify candidate sequence variants in dairy cattle.
Key words: genome-wide association study, dairy cattle, gene mapping

INTRODUCTION

Over a decade ago, genomic prediction was implemented in the routine genetic evaluation of dairy cattle (Hayes et al., 2009; Lund et al., 2011). Genotype data generated from reference and test animals have greatly facilitated GWAS to identify common SNPs that are associated with economic traits. The aim of GWAS is to identify genetic variants that have phenotypic relevance, thus uncovering causal relationships. The GWAS results may enhance the accuracy and persistency of genomic prediction, enable use of increasingly complex models that go beyond additive effects, elucidate the genetic architecture of particular traits, and finally, decipher the underlying biology of complex traits. The entire process, encompassing data generation, quality control, statistical analyses, interpretation of results, and linking findings to biology should be designed and executed to minimize the generation of false-positive and false-negative associations and misleading links to biological processes. False-positive associations may stimulate misdirected studies to identify causal variants, and may thus consume valuable resources and time; or, if prioritized in genomic selection models, may lower prediction accuracy.

Current SNP array genotyping technologies and allele-calling algorithms ensure reliable marker genotypes for GWAS. However, common SNP arrays used in routine genotyping of dairy cattle have a limited number of markers and cannot adequately tag millions of sequence variants. Consequently, to target causal variants, the whole-genome sequence (**WGS**) must be analyzed but only a limited number of animals are sequenced and this panel is neither large enough nor appropriate in terms of phenotype for GWAS. Therefore, a practice that becomes more and more common in dairy cattle is to impute genotypes of WGS variants in the mapping population of animals genotyped with SNP arrays.

Objectives of GWAS are to detect significant associations between genetic variants and the phenotypes for the studied trait by estimating the effects of each variant and then to target the best candidate variants. The most classical method tests each variant individually using linear mixed model, while considering the structure of the population to avoid spurious false-positive associations (e.g., Balding, 2006; Yu et al., 2006). This approach has an important shortcoming: all variants in

Received August 24, 2022.
Accepted February 1, 2023.
*Corresponding author: goutam.sahana@qgg.au.dk

linkage disequilibrium (**LD**) with a causal variant have a statistical effect that can also be significant, which then implies additional work to select only the best candidate variants. More recently, it has been proposed to analyze all variants simultaneously, either by genomic BLUP (**GBLUP**; Gualdrón Duarte et al., 2014) or Bayesian (de Los Campos et al., 2022) methods. These multimarker approaches have the advantage of reducing the effect of long-range LD and to provide narrower location confidence intervals. Because of its ease of use, the GBLUP, or equivalently the SNP-BLUP, is increasingly adopted to perform association analyses in cattle and pigs (Bernal Rubio et al., 2016; Aguilar et al., 2019). It has been shown that statistical tests of linear mixed model and GBLUP based methods are equivalent (e.g., Chen et al., 2017). However, GBLUP and SNP-BLUP based methods have 2 important limitations (see Appendix 1 for details). The first limitation is very strongly regressed SNP effects estimates (i.e., the estimated effect of the top variant is much lower than the true effect). The second limitation is that such approaches are only applicable with a moderate number of SNPs (i.e., a few tens of thousands of markers).

Bayesian methods (Appendix 2) such as BayesR (Erbe et al., 2012), which allow different a priori distributions of SNP effects, and therefore, less shrinkage for the top variants, are very attractive theoretical alternatives but they remain difficult to use at the sequence level because of the very high level of LD between close variants and, more importantly, their huge computation cost. Because the objective of this review is to provide good practices to identify the best candidate causal variants in dairy cattle, we will focus on classical GWAS at WGS level, considering each variant individually.

This review aims to provide a good practice for GWAS at WGS level of quantitative traits in dairy cattle that includes data quality control, association tests, control of population structure, and significance evaluation. We also provide guidance on post-GWAS strategy and interpretation of results. We discuss common limitations and challenges that have been frequently overlooked in the analysis and interpretation of GWAS in dairy cattle. This review is not intended as a detailed guide for performing GWAS in dairy cattle. In addition to the principles, it is meant to provide a good practice so that common mistakes in analysis, interpretation, and reporting can be avoided. For tutorials, protocols, methods and tools, see (Teo, 2008; Weale, 2010; Turner et al., 2011; Reed et al., 2015; Coleman et al., 2016, Ellingson and Fardo, 2016; Marees et al., 2018; Agler et al., 2019; Uffelmann et al., 2021). The common software tools at different stages of genome-wide association analysis and useful databases are presented in Table 1 and Table 2, respectively.

Because no human or animal subjects were used, this analysis did not require approval by an Institutional Animal Care and Use Committee or Institutional Review Board.

THE BASIC STATISTICAL MODEL

Statistical analyses to discover phenotype-genotype associations are described in several previous publications (Balding, 2006; Agler et al., 2019; Wang et al., 2019; Rotroff, 2020; Sesia et al., 2021; Uffelmann et al., 2021). The genetic association between a marker and a phenotype is based on the biometrical model, typically using either a linear model for a continuous phenotype or logistic regression for a binary phenotype. Covariates such as age, sex, and ancestry are included to account for stratification and to avoid confounding by demographic factors and relationship (Uffelmann et al., 2021). Linear mixed model (**LMM**) analysis has become the method of choice when conducting association mapping in the presence of sample structures that may include geographic population structure, family relatedness, or cryptic relatedness (Yu et al., 2006; Kang et al., 2010; Price et al., 2010b). Even if GWAS of dairy cattle are conducted within populations of a single breed, due to large-scale use of AI and limited effective size, the dairy breeds are strongly structured and phenotypes have a marked covariance structure that would affect GWAS results if omitted. In LMM analysis, a polygenic effect is fitted along with genotype for a single SNP as a fixed effect in animal models (Yu et al., 2006; Kang et al., 2008; Sahana et al., 2010). To avoid repeated estimation of environmental effects for each SNP, they are commonly previously estimated and phenotypes are corrected for all systematic nuisance factors. The mixed model then includes only the population mean as the fixed effect beside the SNP effect under investigation. A basic LMM for GWAS can be written as follows:

$$y_i = \mu + bg_i + u_i + e_i, \quad [1]$$

where y_i is the phenotype for the ith individual, μ is the population mean, g_i is the number of copies of the effect allele (0, 1, or 2) of the ith individual, b is the allele substitution effect of the SNP, u_i is the polygenic effect of the ith individual to be normally distributed $N\left(0,\sigma_g^2\mathbf{G}\right)$, where σ_g^2 is the polygenic genetic variance and $\mathbf{G}$ is the additive relationship matrix derived from the pedigree or genome-wide markers; e_i is the random residual for the ith individual to be normally distrib-

Table 1. The common software tools at different stages of genome-wide association analysis

Step	Task	Software tool	Best practice
Phenotype	Corrected phenotype	Basic statistical software	Remove phenotypic outliers Transform the phenotypes when they deviate much from the normal distribution Use de-regressed proof (DRP) or daughter yield deviation (DYD) instead of EBV Adjust the phenotypes for systematic nongenetic effects Compute statistical power for association study based on sample size for a locus explaining 1% of the genetic variance and minor allele frequency (MAF) 10%
Genotype	Quality control (QC)	PLINK (Purcell et al., 2007), VCFtools (Danecek et al., 2011), SNPassoc (González et al., 2007), GWASTools (Gogarten et al., 2012)	Several articles have details on QC (e.g., Marees et al., 2018) Generate sample call and marker call rates Identify sex mismatches and relationship errors (recorded versus genetic), mosaicism, contamination, and sample swaps Discard individual samples and markers that fail QC criteria MAF threshold dependents on the study design and sample size Exclude markers with Hardy-Weinberg equilibrium *P*-value 1.0e-6 Examine presence of population structure
	Mapping the location of SNPs	BLAST (Altschul et al., 1990)	Get uniform genomic location across studies, if SNP positions are not based on same genome assembly
	Phasing	Beagle (Browning et al., 2018), Eagle (Loh et al., 2016b), SHAPEIT (Delaneau et al., 2013)	Pre-phase (estimate haplotypes) before imputation is a good practice, but depends on imputation software used
	Imputation	Minimac4 (Das et al., 2016), IMPUTE2 (Howie et al., 2009), Beagle (Browning et al., 2018), Fimpute (Sargolzaei et al., 2014), AlphaPeel (Whalen et al., 2018)	Apply quality filters to imputed SNPs and exclude based upon any of several available criteria based upon thresholds of imputation quality score, high missing data rate after imputation, or other available metrics Further exclude rare (MAF < pre-set threshold) and monomorphic SNPs
Association analysis	Identify associated markers	GCTA (Yang et al., 2011a), EMMAX (Kang et al., 2010), fastGWA (Jiang et al., 2019b)	GWAS with small sample sizes are prone to type I and type II errors Use linear mixed models (LMM) to test variant-phenotype associations LMM can control type I error rate and increase power in dairy cattle GWAS, but at reasonable computational cost Exclude candidate marker region from the genomic relationship matrix Alternatively, use leave-one-chromosome-out (LOCO) approach of GCTA software Generate QQ plots to check genomic inflation and estimate the lambda to examine residual population stratification In lambda is high (>1.3), use top eigenvectors as covariates and rerun LMM and investigate if inflation is due to some unadjusted structure in the data Use Bonferroni multiple testing corrected *P*-value as the conventional genome-wide statistical significance threshold Visualize the distribution of phenotypes across 3 genotype classes for lead SNP of each associated region
	Meta-analysis	METAL (Willer et al., 2010), MR-MEGA (Mägi et al., 2017)	Use weights proportional to the square root of each study's sample size or by each study estimates' standard errors for meta-analysis When populations or breeds are distant, MR-MEGA, an approach used for trans-ethnic meta-analyses may be more appropriate Meta-analysis hits may be replicated based on directional consistency, nominal or (ideally) genome-wide significant association in external, independent samples For each SNP, report key information (e.g., chromosome position, reference genome build, strand, reference allele, MAF, sample size, effect estimate and its standard error, *P*-value, imputation quality score, as well as meta-analysis statistics) Share or upload GWAS results onto a study-specific or "community" server, to enable harmonization and meta-analyses or dissemination of results
	Multitrait analysis	MTAG (Turley et al., 2018)	Joint analysis of summary statistics from GWAS of different traits, possibly from overlapping samples
	Fine mapping	CAVIAR (Hormozdiari et al., 2014), SuSIE (Wang et al., 2020), FINEMAP (Benner et al., 2016), PAINTOR (Kichaev et al., 2014)	For each associated region, get a credible set of variants those will include the true causal variants with a high confidence

Continued

Table 1 (Continued). The common software tools at different stages of genome-wide association analysis

Step	Task	Software tool	Best practice
Post-GWAS analysis	Variants annotation	VEP (McLaren et al., 2016), ANNOVAR (Wang et al., 2010)	For the credible set of variants, provide annotation for genomic context, relative position or distance from exons or exon boundaries and gene promoter regions Investigate the predicted effect of nonsynonymous coding variants on the protein structure and function using PolyPhen-2
	Regional plot	locusZoom (Pruim et al., 2010) Integrative Genomics Viewer (IGV) (Robinson et al., 2017)	Create regional association plots using LocusZoom Visualize, compare, and contrast results of GWAS findings of the same, similar, or potentially related traits using IGV
	Gene-based analysis	PLINK (Purcell et al., 2007), VEGAS (Mishra and Macgregor, 2015), MAGMA (de Leeuw et al., 2015), PASCAL (Lamparter et al., 2016)	Conduct gene-centric and pathway or gene-set enrichment analyses Examine individual marker associations with functional genome elements and gene expression
	Transcriptome-wide association studies (TWAS)	S-PrediXcan (Barbeira et al., 2018); cGTEX (https://cgtex.roslin.ed.ac.uk; Liu et al., 2022), UTMOST (Hu et al., 2019)	TWAS integrate GWAS and gene expression data sets to identify gene–trait associations
	Annotation enrichment	GARFIELD (Iotchkova et al., 2019)	The interrogation of disease-associated variants' correlation with relevant tissue-specific gene expression is a promising follow-up strategy of GWAS signals
	Partitioned heritability	LDSC (Bulik-Sullivan et al., 2015), LADK (Speed et al., 2020)	Tools for estimating heritability and genetic correlation from GWAS summary statistics

uted $N\left(0,\sigma_e^2\mathbf{I}\right)$, with σ_e^2 the error variance, and $\mathbf{I}$ is an identity matrix. In the next part of this review we will discuss all the details relative to this model.

PHENOTYPE

Sample Size and Power

The GWAS can be applied to any genetically determined trait. Genetic factors can be either major genes or QTL. A GWAS is especially efficient when genetic determinism is simple, with few genes involved, but less adapted to a highly polygenic determinism that may feature many QTL with small effects. Several factors affect the detection power.

Obtaining high-quality phenotypes is the first critical step of GWAS. Quantitative traits in dairy cattle are subject to systematic and random errors, which can influence power and precision. Availability of very large sample sizes in dairy cattle (data from tens or hundreds of thousands of cows) may facilitate the detection of true genetic signals even when phenotypes are not accurate (e.g., the result of one insemination as an indicator of fertility); however, for a given sample size, more accurate definitions and registrations of phenotypes provide more reliable results.

Statistical power, which is the probability to detect a true association, is a function of the true effect (β), i.e., the allele substitution effect (Falconer and Mackay, 1996), and therefore, QTL with larger effects are easier

Table 2. The useful database for cattle GWAS and post-GWAS

Database	Website	Purpose
Animal genome	https://www.animalgenome.org/repository/cattle/UMC_bovine_coordinates	Lift-over between cattle genome build
SNPchiMp v3	https://webserver.ibba.cnr.it/SNPchimp/	Array information
EVA	https://www.ebi.ac.uk/eva/	Variant database
dbSNP	https://www.ncbi.nlm.nih.gov/snp/	Variant database
Animal QTL	https://www.animalgenome.org/cgi-bin/QTLdb/index	Animal QTL database
OMIA	https://omia.org/	Mendelian inheritance in animals
OMIM	https://www.omim.org/	Mendelian inheritance in man
MGI	http://www.informatics.jax.org/vocab/mp_ontology	Mouse mutation lines and phenotype
Ensembl	https://www.ensembl.org/index.html	Central database for animals
GWAS Catalog	https://www.ebi.ac.uk/gwas/	GWAS hits in human
GWAS Atlas	https://atlas.ctglab.nl/	GWAS hits in human
ICAR	https://www.icar.org/	Phenotype definition
FAANG	https://www.faang.org/	Functional annotation for animals
FarmGTEx	https://www.farmgtex.org/	A public resource for regulatory variant discovery and molecular phenotype prediction in farm animal species

to detect. We can assume the effect size estimate $\left(\hat{\beta}\right)$ is normally distributed; $\hat{\beta} \sim N\left[\beta,\sigma^2\left(\hat{\beta}\right)\right]$. Under the null ($\beta = 0$), the t-test statistic $t = \frac{\hat{\beta}}{\sigma(\widehat{\beta})}$ is used to derive the P-value. In the simplest model including only a mean (μ) and a SNP effect (β), the standard error, $\sigma\left(\hat{\beta}\right) = \sigma_e \Big/ \sqrt{2np\left(1-p\right)}$ [because $\mathbf{g}'\mathbf{g} = 2np(1-p)$, where $\mathbf{g}$ is the vector of genotypes]. It decreases as the sample size (n) and the minor allele frequency (**MAF**) of the SNP (p) increase (Spencer et al., 2009; Sham and Purcell, 2014). Power increases when error variance decreases (i.e., when heritability increases) because environmental noise is reduced.

Different strategies can be used to increase the heritability of phenotypes: when available, we recommend the use of the mean of repeated records which has a higher heritability, due to the residual variance divided by the number of records, than each single observation. Another approach considers an intermediate phenotype, for example a particular milk component, a gene expression or a metabolite, with a simpler determinism involving fewer genes. Stronger and more significant gene effects are expected on the intermediate trait than on the complex trait. For instance, variants of the *SLC37A1* gene have a much more significant effect on the phosphorus content of milk, which is associated with caseins in milk, than on milk coagulation, thus allowing more accurate candidate variant targeting (Sanchez et al., 2019).

Of the other parameters affecting power, sample size (n) is under the control of study design. Increasing n decreases standard error in regression models in proportion to $1/\sqrt{n}$. As in any statistical test, underpowered design leads to a proportion of false-negative QTL and an overestimation of the significant effects (Goring et al., 2001), therefore, a high detection power is strongly recommended. To design a GWAS experiment, we have to consider that the largest QTL commonly explain a small proportion of the genetic variance (Hayes and Goddard, 2001) and that most variants have relatively low MAF (Daetwyler et al., 2014). As an example, let us assume a QTL with effect a and MAF p, thus explaining a part of variance equal to $\sigma_s^2 = 2p\left(1-p\right)a^2$ (Falconer and Mackay, 1996). If this variance represents a proportion $prop$ = 1% of the genetic variance, the expected squared t-test is

$$t^2 = \frac{a^2 2p\left(1-p\right)n}{\sigma_e^2} = n \times prop \times \frac{\sigma_g^2}{\sigma_e^2} = n \times prop \times \frac{h^2}{\left(1-h^2\right)},$$

and the sample size n is $n = t^2\left(1-h^2\right)/\left(prop \times h^2\right)$. For a desired t-value of 4 (for instance) and trait with heritability equal to 0.3 and 0.1, the size of the design n would be at least 3,733 and 14,400, respectively. This simple example shows that typical GWAS designs must be large to get high detection power.

Sampling of Individuals

The inclusion of all individuals with the particular phenotype is preferred. However, a focus on population sampling is desirable or even mandatory in several situations. As demonstrated by (Muranty and Goffinet, 1997), when genotyping is the limiting factor in a large population with phenotypes, sampling the animals with extreme phenotypes enhances detection power (compared with random sampling) by increasing the differences in causative allele frequency between groups. However, this procedure strongly overestimates QTL effects because the difference between groups reflects many effects (including error terms) other than the candidate SNP effect. The authors showed that the relative bias is a function of the ratio between the standard deviation in the sample and the standard deviation in the population (Muranty and Goffinet, 1997). Similarly, the proportion of variance explained by the QTL, or heritability in the selected population, is strongly overestimated. A phenotype may be either unexpressed or partially expressed under particular environmental conditions. It is important to record phenotypes in the environments where the phenotype is most expressed. For instance, resistance to disease is expressed only if animals are exposed to the pathogen of interest. Including herds with little or no exposure will underestimate or fail to detect QTL effects. More generally, genetic x environment interactions tend to decrease power as well as relevance of GWAS results.

Another method of population sampling is the case-control design, especially popular in human genetics (Schulz and Grimes, 2002; Zondervan and Cardon, 2007). This is particularly suited to examine rare events. The rationale is to better balance the analyzed phenotypes with the assumption that QTL frequencies are also better balanced. A careful selection of cases and controls allows the homogenization of environmental conditions (e.g., same pathogenic exposure) or genetic background (e.g., same parents), thereby increasing statistical power. The precision of the phenotype (i.e., case/control definitions) should be very clear with explicit inclusion and exclusion criteria. Controls should be derived from the same source population as the cases, and covariate distributions should be similar in both groups.

Definition of the Phenotype to be Included in GWAS

Phenotypes depend on both genetic and environmental effects. A GWAS can use a comprehensive model that includes environmental factors in addition to SNP effects. However, as previously mentioned, frequently used assumptions presume that confounding of SNP and other fixed effects are limited and that phenotypes can be adjusted for these environmental influences without bias on SNP effects. This greatly simplifies the model used for GWAS, which may not include environmental effects to be repeatedly estimated for each SNP. A convenient derivation of corrected phenotypes estimates the effects of environmental factors in a BLUP or GBLUP model and adjusts the phenotype for these effects, or equivalently, considers the sum of the genetic value and residual of the model. In cases of repeated records, as for example cows' milk performances repeated over several parities, the average adjusted phenotype is considered. This new variable is frequently referred to as yield deviation (**YD**; VanRaden and Wiggans, 1991), or more generally as trait deviation.

Not all individual cows with records are genotyped, although their sires usually are. In such a situation, the average performance of the progeny, adjusted for environmental effects and for breeding value of the mates, is a phenotype of choice, and is designated as daughter YD (**DYD**; VanRaden and Wiggans, 1991). The DYD is equivalent to an own performance of the sire, but with a heritability equal to its reliability, which increases with the number of progeny and can therefore be high even when the heritability of the trait is low. This solution is of particular interest when the number of sires is high and, therefore, is restricted to large populations.

The direct use of EBV as phenotypes for GWAS would seem simple and advantageous, because EBV are adjusted for environmental factors. However, EBV are regressed estimates, and their variance depends on their reliability (R^2); that is, on the quantity of available information: $var\left(EBV\right) = R^2\sigma_g^2$, with σ_g^2 the genetic variance. A strongly heterogeneous R^2 may promote spurious associations between SNPs and EBV and increase type-I error (Ekine et al., 2014). Even if reliabilities are similar between individuals, EBV are regressed; therefore, SNP effects would be underestimated. Unless highly reliable, EBV are not recommended as response variables for GWAS.

When only EBV are available, de-regressed EBV, also known as de-regressed proof (**DRP**), are good proxies of YD or DYD. The DRP can be obtained from EBV, reliabilities, and pedigree, and do not require original data for their derivation. The DRP are defined as virtual performances that provide the same EBV when analyzed with a simple model that includes a mean and a genetic effect (possibly with groups). Different methods have been proposed to compute DRP (Garrick et al., 2009; Calus et al., 2016). A DRP is characterized by its weight, $w : var\left(DRP\right) = \sigma_g^2 + \sigma_e^2 \,/\, w$, with σ_e^2 the residual variance.

Weighted Analysis

When the accuracy of (pseudo-)phenotypes varies across individuals, the heterogeneity of residual variance can be accounted for in the model (Garrick et al., 2009). Ignoring this heterogeneity may reduce power and increase the false-positive rate because extreme performances correspond to the most variable error terms. Use of weights is recommended when phenotypes are YD with a variable number of repeated records; for DYD with disparate numbers of progeny; or for DRP with dissimilar proof reliabilities. Weight (w) is the equivalent number of own performances providing the same accuracy. A frequently encountered hindrance is that none of the currently available conventional GWAS software packages implement this heterogeneous variance option. When inclusion of weights is impossible, elimination of the animals with the lowest weights (i.e., those with the least accurate phenotypes) from the analysis may be advisable.

GENOTYPE

Quality Control of SNP Array Genotypes

Rigorous quality control of genotypes before GWAS is essential to optimize accuracy. First, individual-specific quality control of genotype data are undertaken to remove individuals with poor DNA quality or mislabeled samples. Animals with low call rates (generally <95–98%) are excluded from further analysis (Turner et al., 2011). Animals with high or low heterozygosity rates indicate sample contamination or inbreeding, respectively. Individuals deviating ± 3 standard deviations from the mean heterozygosity rate should be excluded (Marees et al., 2018). Heterozygosity levels of SNPs on the X chromosome are used to check inconsistencies in the assigned genetic sex of an individual (Weale, 2010; Turner et al., 2011). A principal component analysis plot of genotyped animals can indicate any outlier samples or detect (sub)clusters in the studied population (Weale, 2010), e.g., large progeny groups (as seen in dairy cattle; Yin and König, 2019). In extensive dairy data sets, genomic relationships can be compared with pedigree relations to detect mislabeled samples or pedigree errors (Calus et al., 2011).

Second, SNP-specific quality control steps are applied to the remaining individuals to remove poorly performing SNPs. Generally, the SNP call rate needs to exceed 95%. The SNPs with genotype frequencies deviating from Hardy-Weinberg equilibrium are excluded. A Hardy-Weinberg equilibrium *P*-value threshold of 10^{-6} is recommended for quantitative traits (Marees et al., 2018), but can vary based on the number of SNPs. Reduction of false-positive SNP-phenotype associations requires a sufficient prevalence of the rare allele in the mapping population. In a population of 1,000 individuals, only 20 copies of the rare allele are expected for a SNP with a MAF of 1%. With such a sample size, SNPs with MAFs below 1% must therefore be removed. The larger the data set, the lower the threshold that can be chosen (Marees et al., 2018). Low MAF SNPs showing significant associations should be checked, because a few animals with extreme phenotypes in low-frequency genotype classes may drive association test statistics.

Quality control of genotypes can be done by using, for example, PLINK (Purcell et al., 2007), VCFtools (Danecek et al., 2011), or specific R packages (https://www.r-project.org/) such as SNPassoc (González et al., 2007) or GWASTools (Gogarten et al., 2012).

Merging Genotype Data

Combining genotype data from multiple sources can be challenging, as generic bovine SNP arrays can be produced by independent manufacturers and can have different densities ranging from 3k to 777k. Furthermore, dissimilar versions may be produced within a density. Custom arrays (or customized parts of a generic array) are designed, produced, and frequently updated by separate manufacturers for use by breeding companies or for use in specific research projects (e.g., EuroGenomics chip; Boichard et al., 2018). The SNP nomenclature may vary between manufacturers; consequently, the matching of SNPs across different chips cannot be based solely on the SNP name.

Chromosomal position should also be constant across arrays when SNP position is based on the same reference genome. When older data are accompanied with a map file based on an older reference genome, a lift-over to the new reference genome must be performed. The University of Missouri-Columbia has created lift-over files for 25 commonly used cattle genotyping arrays to the latest reference genome ARS-UCD1.2 (Rosen et al., 2020) available at https://www.animalgenome.org/repository/cattle/UMC_bovine_coordinates/. SNPchiMp v3, available at https://webserver.ibba.cnr.it/SNPchimp/, is a very convenient program that can combine data from different commercial SNP arrays and can also switch between independent reference genomes (Nicolazzi et al., 2014).

Variants deposited at the NCBI/EVA dbSNP databases (Table 2) are assigned a nonredundant accession number (**rs**). This number is a unique identifier (**ID**), independent of the reference genome. Hence, using these rs-IDs for SNP nomenclature would be highly appropriate, but remains an uncommon practice of array manufacturers. Consequently, finding the correct rs-ID based on SNP name or chromosome position can be tedious; although there are public repositories, not all SNPs may be found.

Each SNP can be uniquely mapped to a reference genome based on its bilateral flanking sequences by using BLAST (Altschul et al., 1990), although such mapping may be time-intensive. These sequences can also be used to match SNPs across arrays and genome builds; however, they are not provided in standard map files.

If common SNPs have been identified across data sets and alleles are coded by the same method, data can be merged. However, depending on the chip, either the forward or reverse strand can be genotyped (e.g., an A/C SNP can become a T/G SNP). Illumina has created a TOP/BOT and A/B allele coding system that will always code the SNP in the same manner across versions and arrays (https://www.illumina.com/documents/products/technotes/technote_topbot.pdf). Genotype data from final reports in all available formats certainly facilitate uniform allele coding. If the same SNP is coded A/C in one file and T/G in another, the T can be changed into A, and G into C for compatibility; however, this is not a trivial task for A/T and G/C SNP, in which case DNA strand information is needed. By design, only a few A/T and G/C SNPs are included on cattle arrays, hence, they may be removed if the necessary information is not at hand.

When combining different batches of populations genotyped with the same or different SNP arrays, comparison of the batch allele frequency across common SNPs can indicate whether the batches can be combined into one study population, and can provide an additional check if allele coding is the same across batches. When plotted, a more or less narrow cluster around the diagonal indicates more or less similarity between the 2 batches. A V- or X-shaped cluster indicates nonuniform allele coding across the batches.

IMPUTATION TO WGS VARIANTS

Reference to Use

Accurate imputation to WGS requires a sequenced reference population with sufficient unique haplotypes,

preferably with a reasonably high number of reference sequences from the same breed to ensure it contains haplotypes specific for the target population (at least 200–300 to have accurate imputation of SNP with MAF of 1%; Druet et al., 2014; Butty et al., 2019). If limited WGS data of the breed are available, a reference combining different breeds can be constructed to enhance imputation accuracy, especially for low MAF variants (Bouwman and Veerkamp, 2014; Korkuć et al., 2019). Because LD at the 50K level is not conserved across breeds, direct imputation from 50K to sequence requires an adequate number of sequenced individuals from the studied breed. When sufficient (several hundred) high-density (**HD**; ~777K SNPs) genotypes are available from the breed, a 2-step imputation from 50K to HD within breed and then from HD to sequence in a multibreed sequenced reference population is usually more accurate. The 1000 Bull Genomes Project was established to ensure a large reference population for imputation to enhance downstream analyses such as GWAS. For all reference individuals, WGS is aligned against the same reference genome; population-wide SNPs and indels (short insertions/deletions) can be called using variant-calling software such as GATK (Van der Auwera et al., 2013) or Freebayes (https://github.com/freebayes/freebayes). This results in genotypes of millions of variants for each individual which are efficiently stored in variant call format (**VCF**) files, a specific format storing large-scale genotype variants, containing meta-data (e.g., reference genome), variant description (e.g., position, alleles), and, possibly, genotypes of individuals (Danecek et al., 2011). To limit errors, it is required to filter variants based on their quality scores or the sequencing depth; guidelines implemented in the 1000 Bull Genomes Project can be followed (Daetwyler et al., 2014). Only bi-allelic variants are generally imputed.

Software and Pipeline

Several software packages can process the phasing and imputation of large WGS data sets. Some software packages manage both phasing and imputation (e.g., Beagle; Browning and Browning, 2007). Others are combined in pipelines that first phase the data with one software package (e.g., Eagle; Loh et al., 2016a) or SHAPEIT (O'Connell et al., 2014) and then impute the data with another program (e.g., Minimac; Fuchsberger et al., 2015; Das et al., 2016) or IMPUTE (Howie et al., 2012). Based on the information source used, imputation programs can be classified into 2 types as follows: (1) population-based, which looks at LD between SNP and haplotype frequencies (e.g., Beagle, IMPUTE, Minimac); and (2) pedigree-based, which looks at cross-generational inheritance of haplotypes [e.g., Fimpute (Sargolzaei et al., 2014) or AlphaPeel (Whalen et al., 2018)]. Software designed for human genetic studies is usually population based, as relatedness in human data sets is limited. Software designed for livestock genetics usually implement both, using pedigree when available and population-based imputation when the pedigree either features low relatedness or is not provided. Several studies have compared accuracies and computational performances of imputation methods (Khatkar et al., 2012; Mulder et al., 2012; Brøndum et al., 2014).

Quality Control After Imputation

Poorly imputed variants should be removed from GWAS to avoid spurious associations. This can be accomplished by using estimated imputation accuracies provided by programs such as Beagle R^2 and Minimac R^2, which are accuracy estimates as the true genotypes are unknown; however, estimated and actual imputation accuracies are highly correlated (van Binsbergen et al., 2014; Pook et al., 2020). Notably, imputation accuracy estimates may strongly vary across imputation software, because of differences in calculation of information metrics at imputed SNPs, and thresholds have to be defined within software (see Supplementary Information S3 in Marchini and Howie, 2010).

The VCF files provide sequence-level genotypes with reference/alternative (**REF/ALT**) allele coding. Usually, array genotypes use different coding (e.g., TOP/BOT for Illumina). As with interchip comparisons, a proper correspondence must be established and array alleles must be coded in REF/ALT before genotypes are mixed.

Dosage Versus Best Guess Genotype Coding

Imputation provides the probability that an individual carries a given genotype. The genotype with the highest probability will be called as the most likely on a discrete scale. Most likely genotypes can be used directly in subsequent analyses and can be applied as true genotypes. Although convenient, using the most likely genotype does not reflect the uncertainty of the imputation algorithm as indicated by genotype probabilities. Hence, GWAS-detected associations may be biased if variants are imputed with low certainty. To consider imputation uncertainty, dosages can be used in an association analysis model. Dosage is the predicted genotype on a continuous scale [i.e., the number of alternative alleles (0, 1, 2) weighed by its probability]. Dosages thus range between 0 and 2, with 0 being certainly homozygous for the reference allele, and 2 being certainly homozygous for the alternative allele.

Not all imputation software packages provide dosages, and some GWAS software can only process integers as genotypes, forcing the use of the most likely genotype. In such cases, filtering of imputation accuracy is more critical.

ASSOCIATION ANALYSIS

Single-Marker Analysis

Suitable (pseudo-)phenotypes for association studies in dairy cattle are described above in the "Phenotype" section. Individuals have multiple levels of familial relatedness, therefore LMM (model 1) is commonly used in dairy cattle GWAS to account for this data structure. A polygenic effect is fitted along with genotype dosage for a single SNP as a fixed effect in animal models (Yu et al., 2006; Kang et al., 2008; Sahana et al., 2010). The mixed model then includes only the population mean as the fixed effect beside the SNP effect under investigation. To account for the differences in accuracies of the (pseudo-)phenotypes, the basic model described earlier (model 1) can be modified by redefining the variance structure of the residual error, $e \sim N\left(0, \sigma_e^2 \mathbf{D}\right)$, where $\mathbf{D}$ is a diagonal matrix equal to identity if all weights are 1, or with diagonal elements $1/w_i$, where w_i is the weight attached to ith record with matrix if phenotype accuracies are different.

Single-SNP analysis using an LMM is computationally intractable for datasets derived from thousands of individuals, owing to the heavy computational burden of estimating variance parameters (Kang et al., 2010). Computation loads can be split into 3 steps: (1) building the genomic relationship matrix [**GRM**; $O(MN^2)$], (2) estimating variance components [$O(MN^2)$], and (3) computing association statistics for each SNP [$O(MN^3)$], where M is the number of markers and N is number of individuals (Yang et al., 2014). In the EMMAX approach (Kang et al., 2010), variance components are estimated only once with a model without SNP effect, and are assumed to be known, with the presumption that the individual SNP contribution to phenotypic variance of a polygenic trait is very small. This is implemented in several software packages, for example EMMAX (Kang et al., 2010) and GCTA (Yang et al., 2011a). However, if one or several large QTL for a phenotype segregating in a population are suggested by an initial GWAS scan, the full LMM can be run to jointly estimate the effects of the major QTL and variance components (e.g. Wu et al., 2016).

Effect size estimates of associated variants are most likely biased upward; this phenomenon is known as winner's curse (Beavis effect; Xu, 2003; Palmer and Pe'er, 2017). Significant association is declared when test statistics reach a predetermined threshold value; therefore, the estimated effects of significant variants are actually sampled from a truncated distribution. Consequently, the effect of an identified variant must be re-estimated from another population (Lande and Thompson, 1990), which is common practice in human GWAS but still too rare in dairy cattle until recently. When possible, we strongly encourage GWAS studies to include confirmation studies.

The GWAS may disclose distinct peaks of significant SNPs, such as QTL regions, but may also reveal isolated significant SNPs. Sporadic significant SNPs are often declared spurious associations; indeed, given the high LD in cattle, it is unlikely that only a single variant in a genomic region picks up the effect of the causal variant. Isolated significant SNPs may also be explained by a low imputation accuracy in the region, false positive or by incorrect positioning of variants in the reference genome (Qanbari et al., 2022).

Conditional and Joint Multiple-SNP Analysis

An important limitation of single-SNP models is their sensitivity to long-range LD, resulting in many correlated significant results. This limitation may be overcome by analyzing SNPs simultaneously.

The most significant SNP from a QTL region is reported as the lead SNP which may not be the causal variant, but is in high LD with the causal variant. The lead SNP may not capture all of QTL variance due to incomplete LD. Alternatively, multiple causal variants could be located within a QTL region. Hence, the variance captured by the lead SNP may underestimate the total variance explained by the region. Conditional analysis is used to identify secondary association signals at a locus involving association analysis conditioning on the primary associated SNP to determine the presence of other significantly associated SNPs (Lango Allen et al., 2010; Cai et al., 2018). Yang et al. (2012) proposed an approximate conditional and joint analysis (**COJO**) using summary-level statistics and LD corrections between SNPs estimated from a reference sample; the method is available on GCTA software (Yang et al., 2011a). COJO can be performed on a data set with individual-level genotype data or by using GWAS summary statistics (e.g., from meta-analysis). The latter requires an individual-level genotype data set representative of the LD pattern in the investigated population. The 1000 Bull Genomes data set can serve as a reference for LD in cattle, preferably including only individuals from the same breed, as LD patterns differ across breeds (de Roos et al., 2008). The SNPs

selected by COJO are assumed to be independently associated with the studied trait, and maximize the captured total variance. The COJO-selected SNPs are a good set for further studies; for instance, they may complement SNPs included in chip arrays for use in genomic prediction models.

Genetic Model

Model of Inheritance. In single-gene association studies as well as in GWAS, the mode of inheritance (additive, dominant, and recessive) is usually not known a priori. Assuming an incorrect mode of inheritance may substantially reduce power, whereas on the other hand, testing all possible models may increase type-I error rates (Bagos, 2013). Exploring the modes of gene action is important to understand the genetic architecture behind quantitative traits. Nonadditive association analysis can be used to scan recessively inherited loci (Reynolds et al., 2021, Reynolds et al., 2022). However, testing for nonadditivity requires enough data in the 3 genotypic classes, especially in the class of homozygotes for the rare allele with expected number equal to $N \times \text{MAF}^2$. Hundreds of thousand individuals are therefore required to test nonadditivity for rare variants. Palmer et al. (2022) tested over 1,000 phenotypes for dominance effects through GWAS scans using UK Biobank data. They estimated that a 20- to 30-fold increase in sample size will be necessary to capture clear evidence of dominance similar to those currently observed for additive effects. Low power was also observed in studies to identify dominance effect in dairy cattle, for example, for milk production traits (Jiang et al., 2019a) and female fertility (Mao et al., 2020; Reynolds et al., 2021). A recessive mode of inheritance is used to identify major genes with deleterious effects, primarily when the phenotypic inheritance indicates a recessive mode of gene action. Notably, to ensure a robust model without an inflated false-positive rate, the dominance effect cannot be tested without the additive effect.

Rare-Variant Mapping. Most genomic variants are very rare (MAF <0.5%) in humans (1000 Genomes Project Consortium, 2015) and rare to less frequent (MAF <5%) in cattle (Daetwyler et al., 2014). Theoretical and empirical studies suggest that rare variants (defined as those with frequencies lower than 1%) may play a significant role in quantitative trait variation (Gibson, 2012; Kemper et al., 2012). However, the mapping of rare variants remains a challenge due to low power and imputation inaccuracy. Rare-variant association studies are generally "gene-based" in the sense that rare variants within the same gene are grouped and then statistically assessed to determine the significance of associations between phenotypes and combined rare variants (Zhang et al., 2016). Guidelines for the combination of rare variants in gene-based analyses were formulated by MacArthur et al. (2014). A common approach collapses rare variants of a gene region into a single meta-allele to represent a genetic burden score (Madsen and Browning, 2009; Price et al., 2010a). Zhang et al. (2016) compared association mapping approaches for rare variants in which samples resemble cattle family structure, and recommended rare-variant specific association mapping methods to overcome confounding of extreme phenotypes in the family mean.

Population Stratification. Large sample sizes are needed to confer adequate power to identify associated loci (Sham and Purcell, 2014). This requirement increases the risk that all samples may not originate from a single homogeneous population. Difference in allele frequencies due to population stratification, cryptic relatedness, and confounding by nongenetic factors may inflate test statistics across the whole genome (Marchini et al., 2004). A formal model of hidden relatedness based on the coalescent theory (Voight and Pritchard, 2005) also suggests a constant inflation across the genome when the sample structure is entirely due to hidden relatedness (Devlin and Roeder, 1999). Populations used in dairy cattle GWAS are constituted of related individuals; consequently, accounting for population structure or genetic relationships between individuals is crucial to avoid inflating the type 1 error rate. Population stratification can be detected by the quantile-quantile plot, which plots observed test statistics against the values that would be obtained from a theoretical distribution under the null hypothesis (Price et al., 2010b). Deviation from the diagonal line indicates possible population stratification and an inflation of spurious associations.

Accounting for Population Structure. One method to detect and correct for population structure is to estimate principal components (**PC**) over the genotype data and to fit them in the model as covariates (Price et al., 2006). As PC reflect ancestry information, they are generally included in the model when multiple populations or breeds with different degrees of relatedness are evaluated in the same GWAS (Zhao et al., 2018). A recommended approach in studies in which the sample size of each population is sufficiently large comprises an initial independent within-breed association analysis followed by the combination of within-breed GWAS results in a meta-analysis (Mao et al., 2016).

In LMM, genetic relatedness between individuals is accounted for by the additive genetic effects of the individuals (i.e., breeding values that are estimated via a covariance matrix describing the genetic relationships between individuals of the population). The relationship matrix can be constructed from pedigree (**PRM**)

or genomic (**GRM**) information. Each approach, equivalent to a BLUP or a GBLUP model, respectively, has advantages and limitations. The PRM based on the expected genetic resemblance between known relatives has computational advantages as its inverse is very sparse and can be obtained directly without computing the matrix itself. Alternatively, GRM derives the proportion of the genome that is identical across animals from SNP genotypes distributed over the whole genome (VanRaden, 2008). It can be obtained even when the pedigree is unknown or incomplete. A medium-density of SNPs (50K) is generally recommended for constructing GRM because the gain in accuracy obtained by using HD SNP panels (777K) is small due to the long-range LD within dairy cattle populations (Su et al., 2012). By estimating realized–and not expected–relationships between individuals, a GRM leads to more accurate breeding values than a PRM (VanRaden, 2008, Hayes et al., 2009); however, the inversion of dense *N*x*N* GRM may consume computational runtime and resources inordinately, or may even prove to be impossible with standard methods when large cohorts are analyzed in GWAS.

Tools have been developed to overcome these computational drawbacks. For example, the fastGWA algorithm, implemented in GCTA (Jiang et al., 2019b), controls for population stratification by adding PC, and for genetic relatedness by sparse GRM (after neglecting the lowest terms; e.g., lower than 0.05). This tool is very efficient for the analysis of large samples of unrelated humans, but is less well-adapted to dairy cattle that are much more related.

The APY approach (Misztal et al., 2014) would be an interesting alternative but is not yet implemented in common GWAS software at sequence level. Alternatively, SNP effects may be estimated by using the SNP-BLUP model instead of estimating breeding values (GBLUP model). Both models are fully equivalent (Strandén and Garrick, 2009). The equation system is large ($M \times M$) but of constant size, and can therefore accommodate very large populations in GWAS. Reynolds et al. (2021) recently applied this approach in a GWAS with more than 130K cows.

Using SNP information to account for the relatedness of individuals can reduce power due to double-fitting the candidate marker in the model, both as a fixed effect tested for association and as a random effect as part of the GRM (Yang et al., 2014). An LMM that excludes candidate markers from the GRM is the mathematically correct approach (Listgarten et al., 2012), but is not implemented in current software. Therefore, this is usually addressed by omitting SNPs on the same chromosome as tested SNPs from the GRM (leave-one-chromosome-out, **LOCO**); however, this approach may inflate test statistics (Yin and König, (2019); (Mesbah-Uddin et al., 2022). Leaving the flanking region of the targeted SNP (leave-one-segment-out, **LOSO**) would be a good practice (e.g., 10 Mb; Reynolds et al., 2021), although Yin and König (2019) suggested a statistical model that considers LOCO plus chromosome-wide PCs.

Correction for Genomic Inflation

Genomic inflation (λ) is measured as the ratio between the median of the observed chi-squared test statistics (which is the squared *t*-statistics) and the expected median of the chi-squared distribution $\left(\chi_1^2\right)$, which is 0.456. The genomic control approach (Devlin and Roeder, 1999; Bacanu et al., 2002), based on the distribution of test statistics from single-marker analysis, is used to estimate the inflation factor. The lambda is used to rescale test statistics to avoid the risk of false positives. The adjusted test statistic $\chi^2_{adj} = \chi^2/\lambda$ follows a chi-squared distribution with 1 df under the null hypothesis. In human GWAS, a $\lambda < 1.1$ is considered lack of evidence for inflation of test statistics (Wellcome Trust Case Control Consortium, 2007). In dairy cattle, higher λ values are expected due to multiple factors. A significant inflation of test statistics is to be expected under polygenic inheritance even in the absence of population structure (Yang et al., 2011b). Yang et al. (2011b) showed that under such conditions, the genomic inflation factor is not expected to be unity, but is a function of sample size, LD structure, number of causal variants, and trait heritability. The spread of LD in dairy cattle breeds is high due to small effective population sizes. The phenotypes used in GWAS are frequently pseudophenotypes with high reliability ("heritability"). Whether high lambda (>1) is due to population structure can be determined by adding a few top PC in the mixed model analysis (Price et al., 2006). Another method to test for population structure is LD score regression analysis (Bulik-Sullivan et al., 2015) using χ^2 statistics of SNPs. In the absence of confounding biases, such as cryptic relatedness and population stratification, the intercept of the LD score regression model would be close to 1.

Multiple Testing Corrections

A typical GWAS conducts hundreds of thousands to millions of tests independently, each for a single marker and with its own false-positive probability. The cumulative likelihood of finding one or more false-positive associations over an entire GWAS is therefore much higher. Because increasing the number of false

discoveries is problematic, GWAS uses methods that control the family-wise error rate (**FWER**) stringently. The FWER is the probability of making at least one false discovery across all tests conducted in the multiple testing setting: $FWER = Prob\ (FD \geq 1)$. In the analysis plan, the multiple testing correction approach must be specified a priori (i.e., before test statistics are observed). The post hoc setting may allow tweaking of multiple testing criteria until significant associations are declared. The inference procedure should fix a statistical significance threshold α, and call each variant significant if its P-value turns out to be $\leq\alpha$. A variety of methods may be used to correct for multiple testing.

Bonferroni Multiple Testing Correction

The most commonly used method to control FWER at a level α is to apply significance threshold $\alpha_c = \alpha/M$ for each test, where M is the number of markers. This is called the Bonferroni correction for multiple testing. However, a frequent criticism of the Bonferroni correction is that it was developed for independent tests, and is therefore inappropriate for GWAS because of LD among markers; therefore, the Bonferroni correction is conservative. Although the method controls false positives rigorously, many true genetic signals are not identified (loss of power). A high significance threshold decreases power and increases bias in the results. The typical European human population has an estimated 1 million independent chromosomal segments. Thus, a P-value threshold of $\frac{0.05}{10^6} = 5\times10^{-8}$ (Pe'er et al., 2008; Fadista et al., 2016) has become a standard for common-variant GWAS in humans (Risch and Merikangas, 1996; Dudbridge and Gusnanto, 2008; Panagiotou and Ioannidis, 2012; Chen et al., 2021). Such a threshold can be applied in dairy cattle GWAS, but it is highly conservative as the number of independent chromosomal segments within a population is directly related to the effective population size, which is generally very low in dairy cattle populations (Doekes et al., 2018; Makanjuola et al., 2020; Gautason et al., 2021).

False Discovery Rate

Calculation of the false discovery rate (**FDR**) is another common approach used in genomic investigations such as studies of gene expression (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). Although FWER-based procedures control the probability of incorrectly rejecting at least one true null hypothesis, FDR-based procedures control the expected proportion of incorrectly rejected true null hypotheses. At a fixed threshold, control of FDR is less conservative than control of FWER (Goeman and Solari, 2014).

The q-value is the FDR-based measure of significance that can be calculated simultaneously for multiple hypothesis testing (Storey, 2002; Storey and Tibshirani, 2003). The q-value is defined as the minimum FDR at which the test is called significant. This approach is based on the observation that if the requirements of the t-test are met, the distribution of its P-values for comparisons for which the null hypothesis is true is expected to be uniform (by definition of the P-value). In contrast, comparisons corresponding to an effect will have more P-values close to zero. The q-value method adjusts P-values with a rank scheme similar to that of Benjamini-Hochberg (1995) but incorporates the estimate of the fraction of tests for which the null is true, π_0 (http://github.com/jdstorey/qvalue).

Based on Effective Number of Independent Tests

Another group of methods extend the Bonferroni or FDR adjustments to account for the correlation between SNPs (i.e., LD among SNPs). A solution is to estimate and use the effective number of independent SNPs and an effective number (M_{eff}) of independent tests (Cheverud, 2001; Li and Ji, 2005; Hendricks et al., 2014). This approach is considered less conservative than the Bonferroni correction and more computationally efficient than permutation testing; however, its accuracy was unsatisfactory when applied to GWAS (Dudbridge and Koeleman, 2004; Salyakina et al., 2005). Its use in dairy cattle requires careful consideration due to a high range of LD.

The P-value quantifies incompatibility between observed data and the null hypothesis, and is generally used as a label to declare statistical significance. The P-value threshold should be fixed before test statistics are observed. When testing a subset of markers, e.g., one genomic region or one chromosome, the P-value threshold should not be relaxed compared with what is set for the whole-genome level. Furthermore, exact P-values are more informative than binary significant or nonsignificant classifications.

Meta-Analysis

A GWAS meta-analysis can improve both the power of discovery and the accuracy of mapping. This approach combines summary-level results from K independent GWAS (GWAS utilizing nonoverlapping samples) on the same phenotype into a single estimate (Evangelou and Ioannidis, 2013). Hence, for each variant, the issue is how to combine K independent GWAS

association statistics into a single estimate of the effect size, standard error, and *P*-value. Meta-analysis at the WGS level is recommended when GWAS results come from different populations or breeds. Statistical power and mapping resolution are expected to improve for causal mutations shared among the different populations or breeds. Indeed, power is increased because a meta-analysis corresponds to a larger design than each of the individual GWAS; a better resolution is expected because LD is more limited in the whole meta-population than within each population. The sharing of primary data is a limiting factor for joint analysis of dairy cattle, as data are often owned by commercial breeding companies. Furthermore, multiple sources of heterogeneities may preclude joint analysis of primary data. Although meta-analysis obviates the sharing of sensitive individual-level genotype-phenotype data, it also restricts the set of possible downstream analyses, because only marginal additive effect estimates are available. To maximize the scientific output of consortia efforts, meta-analyses should be designed so that all consortium members follow identical data quality control and analytical procedures before sharing summary statistics. The fixed-effects model is commonly used, and assumes that all studies measure the same underlying quantity (i.e., effect size is the same across all studies). The z-score approach, which converts the *P*-value and the effect direction in z-score [$z = \Phi^{-1}(1 - P/2) \times$ effect sign] without considering the magnitude of the effect, may be a good alternative when the traits measured in the different studies are not identical. Both approaches weight different studies by their sample sizes and are implemented in METAL software (Willer et al., 2010). A few check points are (1) correct naming of files so that they may be traced back to a specific cohort, (2) flipped alleles, (3) duplicated SNPs, (4) bad imputation quality, and (5) incorrect analysis models to account for population stratification and relatedness among individuals. A large meta-analysis of cattle stature was conducted by an international consortium (Bouwman et al., 2018). Bernal Rubio et al. (2016) presented meta-analysis of genome-wide association from genomic prediction models.

The standard approach of genetic association studies is to analyze a single trait. Such studies do not exploit information contained in summary statistics from GWAS of related traits. Multitrait analysis of GWAS enables joint analysis of multiple traits, thus boosting statistical power to detect the genetic associations of each trait (Turley et al., 2018); subsequent meta-analyses of summary statistics are used to identify loci with pleiotropic effects. Bolormaa et al. (2014) used each SNP effect on each trait derived from single-trait GWAS and genetic correlations between traits to estimate test statistics for multitrait meta-analysis. They estimated genetic correlations between traits as overall correlations of all SNP effects (signed t-values) of the 2 traits. However, LD score regression (Bulik-Sullivan et al., 2015) may better estimate genetic correlations from GWAS summary statistics.

FINE MAPPING

Statistical Approaches to Fine Mapping and Visualizing Candidate Regions

The GWAS results do not provide direct insights into causal relationships and underlying biological mechanisms. The GWAS cannot distinguish between causal variants from markers in very high LD with causal variants. In addition, the assumption that a single variant per locus contributes to a trait may be incorrect, as 2 or more variants may exhibit additive or epistatic interactions. Schaid et al. (2018) have reviewed strengths and weaknesses of various statistical approaches for fine mapping. After the initial genome scan, associated SNPs should be partitioned into independent associated regions.

For studies using WGS variant sets, discriminating causal from significant SNPs within association peaks is not trivial. The long range of LD in cattle usually yields peaks of GWAS with many significant SNPs. The top SNPs are often not causal (Cai et al., 2019). Consequently, many fine mapping methods to account for LD have been proposed recently. The CAVIAR method takes an arbitrary number of causal variants when estimating the posterior probability of a variant being causal (Hormozdiari et al., 2016). Another approach, named SuSIE, is based on iterative Bayesian stepwise selection and may account for LD (Wang et al., 2020). FINEMAP software uses GWAS summary statistics to calculate effect sizes and heritability of likely causal SNPs (Benner et al., 2016). FINEMAP can also use LD matrices to account for LD in the mapping population. The PAINTOR uses a probabilistic framework that integrates association strength with functional genomic annotation data to improve accuracy in selecting plausible causal variants (Kichaev et al., 2014).

LocusZoom (Pruim et al., 2010) provides an informative method to visualize GWAS hits. Lead SNPs, surrounding genes, additional tracks of genomic features, LD patterns, and *P*-values can be presented together in a single plot. The Integrative Genomics Viewer (Robinson et al., 2017) is a tool to inspect, validate, and interpret WGS data sets, as well as other types of genomic data and can be used to visualize, compare and contrast results of GWAS findings of the same, similar or potentially related traits.

Replication Study Using Custom SNP Array

A common practice in human GWAS is validating associations with independent samples. Some recent GWAS in dairy cattle have included validation (e.g., Bouwman et al., 2018; Tribout et al., 2020; van den Berg et al., 2020) but this step is often missing in livestock studies due to difficulties in finding unrelated populations, in accessing data or summary statistics, or in funding such research activities. An independent set of samples for QTL validation is generally unavailable from the same discovery population of cattle, but can be validated in other populations of the same or other breeds. Cross-population/breed analysis is advantageous for fine mapping causal variants. Due to the breakdown of LD patterns, noncausal variants will not exhibit the same effect estimates across populations. This will allow differentiation of causal from noncausal variants. For discovery GWAS utilizing bull data, cow data from the same population but not used for bull pseudophenotype calculations can be employed for QTL validation. Factors that may preclude QTL validation include lack of power, cross-population differences in allele frequencies, genetic heterogeneity, and winner's curse.

The addition of candidate variants to customized array used in routine genotyping can generate large numbers of genotypes in diverse breeds efficiently (Boichard et al., 2018). Furthermore, SNP array genotypes are more accurate than genotypes called from sequence data, especially for rare variants. However, most animals genotyped for genomic prediction are young and cannot be used immediately for confirmation studies as the phenotype may be unexpressed due to immaturity. Nonetheless, these animals' genotypes can be used to impute older animals with phenotypes; moreover, imputation is much more accurate than from sequences because of the large number of genotyped animals. Several candidate variants were validated by this approach in French cattle breeds (Michot et al., 2016; Mesbah-Uddin et al., 2019; Tribout et al., 2020).

ASSOCIATION TO BIOLOGICAL FUNCTION

Starting from the Human Genome Project (Sawicki et al., 1993) and next-generation sequencing (Reis-Filho, 2009), the constantly growing body of genomics information has accelerated the exploration of the biological determinates underlying GWAS hits. Associated variants can be prioritized based on their annotation and location in genome-like coding variants in either genes or in regulatory variants of promoters and enhancers. Inferring variants responsible for allele-specific expression, as measured in RNA-seq experiments, can then directly indicate the functionality of particular variants on open chromatin regions. Similarly, variants that disrupt underlying transcription factor binding sites are important candidates for causal variants.

Confirmation in QTL, GWAS, and Disease Databases

The Animal QTL Database (Hu et al., 2022) can provide a starting point to determine if detected QTL regions confirm earlier findings curated in a particular database. In December 2021, the cattle QTL database contained 177,199 QTL from 1,090 publications, representing 689 traits. QTL regions can be prioritized by studying the genes underlying the regions or located close to the QTL. The GWAS results in other species, available in the GWAS Catalog (MacArthur et al., 2017) and GWAS Atlas (Tian et al., 2020), enable the determination of overlap between identified QTL or genes and those reported to encode similar traits. Genes with functions known to be related to the studied phenotype, especially genes included in the Online Mendelian Inheritance in Animals (https://omia.org/), Online Mendelian Inheritance in Man (https://www.omim.org/), or Mouse Genome Informatics (http://www.informatics.jax.org/; Bult et al., 2019) databases can be declared as candidate genes.

Functional Annotation and Genomic Features

The GWAS results are interpreted through the annotation of variants within association signals. The Ensembl variant effect predictor (**VEP**) is a commonly used tool and database for variant annotation (McLaren et al., 2016). By processing a simple input of variant location and nucleotide changes, VEP will output genes and transcripts that harbor the variants, describe the types of variants, predict their consequences, and provide conservation scores of missense variants (i.e., mutations that alter the amino acid composition of proteins). The VEP conservation score is built under the assumption that important AA sequences will be conserved in a protein family and across species; consequently, changes of conserved AA sequences should affect protein function (Ng and Henikoff, 2002).

The variants in regulatory regions are poorly annotated by VEP. An alternative strategy can be to check if candidate variants alter transcription factor binding sites. This can be achieved with multispecies transcription factor binding sites model databases, such as JASPAR (JASPAR CORE 2018 collection; Sandelin et al., 2004), HOCOMOCO (version v10; Khamis et al., 2018), or TRANSFAC (version v3.2 public; Knüppel et al., 1994). An application on dairy cattle is presented in Sanchez et al. (2021).

Genomic features indicate genomic regions that have identified functions. Genomic features include genes (including 5′ UTR, 3′ UTR, intron and exon), rRNA, tRNA, noncoding RNA, pseudogenes, repeat sequences, and regulatory elements. The annotation of protein-encoding genes, rRNA, and tRNA are well defined. Increasing attention is paid to noncoding RNAs, which are important gene expression pathway regulators (Mattick and Makunin, 2006), though, the repeat sequences are not well characterized (Tarailo-Graovac and Chen, 2009). The Gene Ontology Consortium (http://geneontology.org/; Ashburner et al., 2000) and the Kyoto Encyclopedia of Genes and Genomes (https://www.genome.jp/kegg/; Kanehisa and Goto, 2000) provide resources for the identification of the functions of mapped genes. Similar to the ENCODE Project for humans (Encode Project Consortium, 2004), the mapping of potential regulatory elements of the cattle genome has recently become available. Recent discoveries include epigenetic histone modifications of H3K4me3, H3K27ac, H3K4me1, H3K27me3 and the CTCF DNA-binding protein identified by utilizing chromatin accessibility using DNase I hypersensitive site sequencing and the Assay for Transposase-Accessible Chromatin using sequencing in 8 tissues (liver, lung, spleen, skeletal muscle, subcutaneous adipose, cerebellum, cerebral cortex, and hypothalamus; Kern et al., 2021).

Gene-Based and Gene-Set Analyses

Gene-based and gene-set analyses are commonly used to study genes and genome features. These combine the statistics of variants into genes or genomic features and gene sets. To determine the significance of a gene or genomic feature with multiple loci, permutation testing offers an exact assessment that requires few assumptions regarding the computed distribution. Meanwhile, permutation can also account for LD, gene size, and other possible confounders (Liu et al., 2010). However, the computational cost of permutation for GWAS limits its application. The method is implemented in PLINK (Purcell et al., 2007) as a set-based test. As an alternative to permutation, VEGAS discloses the associations of a gene or genomic feature by summarizing the full set of variants (or a defined subset of the most significant variants) in the gene or genomic feature (Liu et al., 2010). This method could be biased due to LD. To overcome LD and the computational demands of permutation, MAGMA uses several methods to test the gene or genomic feature's significance (de Leeuw et al., 2015). To reduce computational demand, MAGMA uses either PC of SNP matrices to prune away SNPs presenting small eigenvalues; or adaptive permutation, a procedure that can vary the number of permutations per gene, depending on its P-value. The construction of a pairwise SNP-by-SNP correlation matrix as implemented in PASCAL represents an alternative approach (Alonso-González et al., 2019).

Functional Analysis

Analyses of databases that include new information or newly designed experiments may facilitate the prioritization of variant or gene candidates. The most straightforward method is to detect eQTL (i.e., genomic regions that affect gene expression levels). There are 2 advantages of combining eQTL with GWAS results from the same population: (1) confirmation of causal variant segregation, and (2) without any prior knowledge, eQTL can map variants affecting gene expression by any possible mechanism (Gilad et al., 2008). This application can also be extended to summary statistics and to individuals for GWAS mapping and eQTL mapping in the same population (Zhu et al., 2016). The combination of GWAS and eQTL is further formalized as a transcriptome-wide association study (Gamazon et al., 2015) that integrates individual-level genotype data or GWAS summary statistics with externally estimated eQTLs. However, major drawbacks of eQTL include the need to select the appropriate tissue, and costs of sampling and sequencing. The challenge of choosing and sampling a suitable tissue may be partially abrogated by using blood, although with some loss of tissue-specific eQTL (Qi et al., 2018). This field of research is expanding very fast in bovine and the recent FarmGTEx database (Liu et al., 2022) compiles the main results on gene expression in cattle.

Because regulatory elements of cattle have been mapped only recently, variant effects on gene expression are still poorly understood. Hi-C (Lieberman-Aiden et al., 2009), based on chromosome conformation capture, can detect comprehensive nuclear chromatin interactions (Belton et al., 2012). Consequently, the interacted genomic region may be mapped to reveal regulatory relationships of GWAS hits.

Enrichment analysis may relate association test results to the regulatory elements of multiple tissue or cell types. GARFIELD can leverage GWAS findings with functional annotations to find features that are relevant to a phenotype of interest (Iotchkova et al., 2019). GARFIELD uses logistic regression to derive the statistical significance of the tested genomic feature while accounting for LD and genomic feature density. In addition, partitioning the heritability of genomic features may also elucidate the contribution of certain types of genomic features to phenotype. Such application may be conducted using LDSC (Bulik-Sullivan et al., 2015) or LADK (Speed et al., 2020). However, both

strategies require well-defined annotations of genomic features to interpret GWAS results.

SHARING DATA AND SUMMARY STATISTICS

A common practice in human GWAS is to deposit summary statistics in publicly accessible databases such as GWAS Catalog (MacArthur et al., 2017) and GWAS Atlas (Tian et al., 2020) to facilitate validation and cross-study joint analyses. However, this approach may be impeded in livestock research because of proprietary interests of breeding companies. Consequently, the advantages of sharing data and information within the breeding industry must be highlighted. Such cooperation will expand access to large population maps of phenotype and genotype, because these data are routinely recorded for management and breeding purposes. GWAS findings should be published in detail to facilitate future research and implementation. Therefore, we discuss in the following paragraphs what type of reporting could be informative in such a setup.

The phenotypes and genotypes of dairy cattle are often owned by breeding companies or associations, and hence cannot be shared easily. In contrast, GWAS summary statistics may be more accessible. The GWAS summary statistics typically include chromosome, position, effect allele, effect size, its standard error, and *P*-value. The minimal requirements to match SNPs across different data sets include chromosome, position, and the reference genome. Reporting the allele on which the effect was estimated (the counted allele) is required to ensure that it was identified as the same effect allele across data sets. Generally, information regarding GWAS accuracy is not shared. Hence, we advise sharing additional SNP information by including imputation R^2 (and imputation software), allele frequency of the SNP in the studied population, as well as phenotypic information including trait heritability, genetic variance, trait definition, and units. Reporting the imputation R^2 and allele frequency can aid the setting of appropriate thresholds across all shared populations. In addition, the size of the studied population should be considered in meta-analysis. Unfiltered GWAS summary statistics should be shared.

The GWAS-derived variant effect and standard error of the effect need to be standardized per population by dividing them by the standard deviation of the phenotypes used in GWAS. The effect size is another parameter to filter spurious SNP associations. If the effect is more than 3 phenotype standard deviations from the mean, the SNP should be excluded from meta-analysis.

Traits that are defined very differently than in the rest of the populations should not be included. Most traits of dairy cattle have been defined by ICAR (https://www.icar.org/index.php/icar-recording-guidelines/), and are available in most countries to ensure consistent trait definitions. However, recently recorded traits such as feed efficiency and methane emission are not yet well standardized, and hence require more elaborate discussions beforehand, or more careful inclusion afterward.

FUTURE CONSIDERATIONS

Methodologies and software for large-scale GWAS are developed primarily by scientists working with human data. There are distinct differences between human and livestock studies; samples are less related in human studies, whereas dairy cattle are characterized by large half-sib families. The LD in dairy cattle is spread over a longer distance due to small effective population size, emphasizing the need for large designs, use of meta-analyses across populations, and confirmation studies. Owing to routine genotyping for genomic prediction, hundreds of thousands of cows are genotyped; consequently, their data could be used for GWAS. However, computer capacities for existing LMM-based tools are exceeded, especially if the number of traits is large. Alternative approaches to processing large data have been suggested (Jiang et al., 2019b), but are suboptimal for dairy cattle. The GRM in humans can set a large part of the relationship to zero, which is not possible in dairy cattle (results in high inflation, lambda value). Therefore, there is a need for easy-to-use GWAS software suited to very large dairy cattle populations. More work is also needed to map genes with nonadditive effect size and to identify pleiotropy. Additional studies of gene-based and rare variants in dairy cattle are required. High LD can also create issues for gene-based mapping methods in dairy cattle. Functional annotation of the cattle genome has nowhere near the quality of analogous human data. The Functional Annotation of Animal Genomes Project (https://www.faang.org/; Andersson et al., 2015) can fill the gap.

The GWAS is funded indirectly through genomic prediction studies. In the future, genomic selection will be based on causal variants that persist over generations and across populations. International efforts and data sharing are necessary to map genes. An understanding of newly designated phenotypes such as feed efficiency, methane emission, and welfare-related traits can also benefit from the mapping of variants. However, phenotype sample sizes are limited, which necessitates the expansion of international collaboration.

CONCLUSIONS

The GWAS of dairy cattle should adhere to analytical rigor to control for multiple testing correction, popula-

tion stratification, and familial relatedness; and should also include appropriate quality controls for both genotype and phenotype data. If the study objective is gene discovery, minimizing the false-positive rate is more important than controlling the false-negative rate. A genome-wide significance threshold (nominal P-value 5×10^{-8} for GWAS) is recommended for GWAS with WGS variants. Replication studies that consider bulls and cows as independent populations should be completed and followed by meta-analysis. If the study objective is to investigate genetic architecture (e.g., to understand the extent to which epistatic or gene–environment interactions control phenotypic variation), controlling for both type-I and type-II errors is extremely important to ensure statistical robustness. If the objective is to validate candidate SNPs by constructing custom SNP arrays, the significance threshold may be relaxed and more weight can be placed on functional annotation of variants.

REFERENCES

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393.

Agler, C. S., D. Shungin, A. G. Ferreira Zandoná, P. Schmadeke, P. V. Basta, J. Luo, J. Cantrell, T. D. Pahel Jr., B. D. Meyer, J. R. Shaffer, A. S. Schaefer, K. E. North, and K. Divaris. 2019. Protocols, methods, and tools for genome-wide association studies (GWAS) of dental traits. Methods Mol. Biol. 1922:493–509. https://doi.org/10.1007/978-1-4939-9012-2_38.

Aguilar, I., A. Legarra, F. Cardoso, Y. Masuda, D. Lourenco, and I. Misztal. 2019. Frequentist P-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. Genet. Sel. Evol. 51:28. https://doi.org/10.1186/s12711-019-0469-3.

Alonso-Gonzalez, A., M. Calaza, C. Rodriguez-Fontenla, and A. Carracedo. 2019. Novel gene-based analysis of ASD GWAS: Insight into the biological role of associated genes. Front. Genet. 10:733. https://doi.org/10.3389/fgene.2019.00733.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Andersson, L., A. L. Archibald, C. D. Bottema, R. Brauning, S. C. Burgess, D. W. Burt, E. Casas, H. H. Cheng, L. Clarke, C. Couldrey, B. P. Dalrymple, C. G. Elsik, S. Foissac, E. Giuffra, M. A. Groenen, B. J. Hayes, L. S. Huang, H. Khatib, J. W. Kijas, H. Kim, J. K. Lunney, F. M. McCarthy, J. C. McEwan, S. Moore, B. Nanduri, C. Notredame, Y. Palti, G. S. Plastow, J. M. Reecy, G. A. Rohrer, E. Sarropoulou, C. J. Schmidt, J. Silverstein, R. L. Tellam, M. Tixier-Boichard, G. Tosser-Klopp, C. K. Tuggle, J. Vilkki, S. N. White, S. Zhao, H. Zhou, and FAANG Consortium. 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. 16:57. https://doi.org/10.1186/s13059-015-0622-4.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: Tool for the unification of biology. Nat. Genet. 25:25–29. https://doi.org/10.1038/75556.

Bacanu, S. A., B. Devlin, and K. Roeder. 2002. Association studies for quantitative traits in structured populations. Genet. Epidemiol. 22:78–93. https://doi.org/10.1002/gepi.1045.

Bagos, P. G. 2013. Genetic model selection in genome-wide association studies: Robust methods and the use of meta-analysis. Stat. Appl. Genet. Mol. Biol. 12:285–308. https://doi.org/10.1515/sagmb-2012-0016.

Balding, D. J. 2006. A tutorial on statistical methods for population association studies. Nat. Rev. Genet. 7:781–791. https://doi.org/10.1038/nrg1916.

Barbeira, A. N., S. P. Dickinson, R. Bonazzola, J. Zheng, H. E. Wheeler, J. M. Torres, E. S. Torstenson, K. P. Shah, T. Garcia, T. L. Edwards, E. A. Stahl, L. M. Huckins, GTEx Consortium. D. L. Nicolae, N. J. Cox, and H. K. Im. 2018. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun. 9:1825. https://doi.org/10.1038/s41467-018-03621-1.

Belton, J.-M., R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker. 2012. Hi–C: A comprehensive technique to capture the conformation of genomes. Methods 58:268–276. https://doi.org/10.1016/j.ymeth.2012.05.001.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate—A practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Benjamini, Y., and D. Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29:1165–1188. https://doi.org/10.1214/aos/1013699998.

Benner, C., C. C. A. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, and M. Pirinen. 2016. FINEMAP: Efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32:1493–1501. https://doi.org/10.1093/bioinformatics/btw018.

Bernal Rubio, Y. L., J. L. Gualdrón Duarte, R. O. Bates, C. W. Ernst, D. Nonneman, G. A. Rohrer, A. King, S. D. Shackelford, T. L. Wheeler, R. J. Cantet, and J. P. Steibel. 2016. Meta-analysis of genome-wide association from genomic prediction models. Anim. Genet. 47:36–48. https://doi.org/10.1111/age.12378.

Boichard, D., M. Boussaha, A. Capitan, D. Rocha, C. Hoze, M.-P. Sanchez, T. Tribout, R. Letaief, P. Croiseau, C. Grohs, W. Li, C. Harland, C. Charlier, M. S. Lund, G. Sahana, M. Georges, S. Barbier, W. Coppieters, S. Fritz, and B. Guldbrandtsen. 2018. Experience from large scale use of the EuroGenomics custom SNP chip in cattle. Page 675 in Proceedings of the World Congress on Genetics Applied to Livestock Production. Molecular Genetics 4. http://www.wcgalp.org/system/files/proceedings/2018/experience-large-scale-use-eurogenomics-custom-snp-chip-cattle.pdf.

Bolormaa, S., J. E. Pryce, A. Reverter, Y. Zhang, W. Barendse, K. Kemper, B. Tier, K. Savin, B. J. Hayes, and M. E. Goddard. 2014. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. PLoS Genet. 10:e1004198. https://doi.org/10.1371/journal.pgen.1004198.

Bouwman, A. C., H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei, F. S. Schenkel, G. Sahana, A. Govignon-Gion, S. Boitard, M. Dolezal, H. Pausch, R. F. Brondum, P. J. Bowman, B. Thomsen, B. Guldbrandtsen, M. S. Lund, B. Servin, D. J. Garrick, J. Reecy, J. Vilkki, A. Bagnato, M. Wang, J. L. Hoff, R. D. Schnabel, J. F. Taylor, A. A. E. Vinkhuyzen, F. Panitz, C. Bendixen, L. E. Holm, B. Gredler, C. Hoze, M. Boussaha, M. P. Sanchez, D. Rocha, A. Capitan, T. Tribout, A. Barbat, P. Croiseau, C. Drogemuller, V. Jagannathan, C. Vander Jagt, J. J. Crowley, A. Bieber, D. C. Purfield, D. P. Berry, R. Emmerling, K. U. Gotz, M. Frischknecht, I. Russ, J. Solkner, C. P. Van Tassell, R. Fries, P. Stothard, R. F. Veerkamp, D. Boichard, M. E. Goddard, and B. J. Hayes. 2018. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat. Genet. 50:362–367. https://doi.org/10.1038/s41588-018-0056-5.

Bouwman, A. C., and R. F. Veerkamp. 2014. Consequences of splitting whole-genome sequencing effort over multiple breeds on im-

putation accuracy. BMC Genet. 15:105. https://doi.org/10.1186/s12863-014-0105-8.
Brøndum, R. F., B. Guldbrandtsen, G. Sahana, M. S. Lund, and G. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15:728. https://doi.org/10.1186/1471-2164-15-728.
Browning, B. L., Y. Zhou, and S. R. Browning. 2018. A one-penny imputed genome from next-generation reference panels. Am. J. Hum. Genet. 103:338–348. https://doi.org/10.1016/j.ajhg.2018.07.015.
Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097. https://doi.org/10.1086/521987.
Bulik-Sullivan, B. K., P. R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale. 2015. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. 47:291–295. https://doi.org/10.1038/ng.3211.
Bult, C. J., J. A. Blake, C. L. Smith, J. A. Kadin, and J. E. Richardson.The Mouse Genome Database Group. 2019. Mouse Genome Database (MGD) 2019. Nucleic Acids Res. 47(D1):D801–D806. https://doi.org/10.1093/nar/gky1056.
Butty, A. M., M. Sargolzaei, F. Miglior, P. Stothard, F. S. Schenkel, B. Gredler-Grandl, and C. F. Baes. 2019. Optimizing selection of the reference population for genotype imputation from array to sequence variants. Front. Genet. 10:510. https://doi.org/10.3389/fgene.2019.00510.
Cai, Z., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2018. Prioritizing candidate genes post-GWAS using multiple sources of data for mastitis resistance in dairy cattle. BMC Genomics 19:656. https://doi.org/10.1186/s12864-018-5050-x.
Cai, Z., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2019. Dissecting closely linked association signals in combination with the mammalian phenotype database can identify candidate genes in dairy cattle. BMC Genet. 20:15. https://doi.org/10.1186/s12863-019-0717-0.
Calus, M. P., H. A. Mulder, and J. W. Bastiaansen. 2011. Identification of Mendelian inconsistencies between SNP and pedigree information of sibs. Genetics, selection, evolution. Genet. Sel. Evol. 43:34.
Calus, M. P. L., J. Vandenplas, J. ten Napel, and R. F. Veerkamp. 2016. Validation of simultaneous deregression of cow and bull breeding values and derivation of appropriate weights. J. Dairy Sci. 99:6403–6419. https://doi.org/10.3168/jds.2016-11028.
Chen, C., J. P. Steibel, and R. J. Tempelman. 2017. Genome-wide association analyses based on broadly different specifications for prior distributions, genomic windows, and estimation methods. Genetics 206:1791–1806. https://doi.org/10.1534/genetics.117.202259.
Chen, Z., M. Boehnke, X. Wen, and B. Mukherjee. 2021. Revisiting the genome-wide significance threshold for common variant GWAS. G3 Genes|Genomes|Genetics 11(2).
Cheverud, J. M. 2001. A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87:52–58. https://doi.org/10.1046/j.1365-2540.2001.00901.x.
Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genetics, selection, evolution. Genet. Sel. Evol. 42:2.
Coleman, J. R., J. Euesden, H. Patel, A. A. Folarin, S. Newhouse, and G. Breen. 2016. Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray. Brief. Funct. Genomics 15:298–304. https://doi.org/10.1093/bfgp/elv037.
Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brondum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858–865. https://doi.org/10.1038/ng.3034.
Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330.
Das, S., L. Forer, S. Schonherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P. R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, and C. Fuchsberger. 2016. Next-generation genotype imputation service and methods. Nat. Genet. 48:1284–1287. https://doi.org/10.1038/ng.3656.
de Leeuw, C. A., J. M. Mooij, T. Heskes, and D. Posthuma. 2015. MAGMA: generalized gene-set analysis of GWAS data. PLOS Comput. Biol. 11:e1004219. https://doi.org/10.1371/journal.pcbi.1004219.
de Los Campos, G., A. Grueneberg, S. Funkhouser, P. Pérez-Rodríguez, and A. Samaddar. 2022. Fine mapping and accurate prediction of complex traits using Bayesian Variable Selection models applied to biobank-size data. Eur. J. Hum. Genet. https://doi.org/10.1038/s41431-022-01135-5.
de Roos, A. P., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179:1503–1512. https://doi.org/10.1534/genetics.107.084301.
Delaneau, O., B. Howie, A. Cox, J. Zagury, and J. Marchini. 2013. Haplotype estimation using sequencing reads. Am. J. Hum. Genet. 93:687–696. https://doi.org/10.1016/j.ajhg.2013.09.002.
Devlin, B., and K. Roeder. 1999. Genomic control for association studies. Biometrics 55:997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x.
Doekes, H. P., R. F. Veerkamp, P. Bijma, S. J. Hiemstra, and J. J. Windig. 2018. Trends in genome-wide and region-specific genetic diversity in the Dutch-Flemish Holstein-Friesian breeding program from 1986 to 2015. Genet. Sel. Evol. 50:15.
Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112:39–47. https://doi.org/10.1038/hdy.2013.13.
Dudbridge, F., and A. Gusnanto. 2008. Estimation of significance thresholds for genomewide association scans. Genet. Epidemiol. 32:227–234. https://doi.org/10.1002/gepi.20297.
Dudbridge, F., and B. P. C. Koeleman. 2004. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am. J. Hum. Genet. 75:424–435. https://doi.org/10.1086/423738.
Ekine, C. C., S. J. Rowe, S. C. Bishop, and D. J. de Koning. 2014. Why breeding values estimated using familial data should not be used for genome-wide association studies. G3 (Bethesda) 4:341–347. https://doi.org/10.1534/g3.113.008706.
Ellingson, S., and D. Fardo. 2016. Automated quality control for genome wide association studies. F1000Res. 5:1889.
Encode Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–640. https://doi.org/10.1126/science.1105136. PubMed
Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114–4129. https://doi.org/10.3168/jds.2011-5019.
Evangelou, E., and J. P. Ioannidis. 2013. Meta-analysis methods for genome-wide association studies and beyond. Nat. Rev. Genet. 14:379–389. https://doi.org/10.1038/nrg3472.
Fadista, J., A. K. Manning, J. C. Florez, and L. Groop. 2016. The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur. J. Hum. Genet. 24:1202–1205. https://doi.org/10.1038/ejhg.2015.269.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to quantitative genetics. 4. ed. Longman, Harlow.

Fernando, R. L., J. C. M. Dekkers, and D. J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet. Sel. Evol. 46:50. https://doi.org/10.1186/1297-9686-46-50.

Fernando, R. L., and D. Garrick. 2013. Bayesian methods applied to GWAS. Pages 237–274 in Genome-Wide Association Studies and Genomic Prediction. C. Gondro, J. van der Werf, and B. Hayes, ed. Humana Press.

Fuchsberger, C., G. R. Abecasis, and D. A. Hinds. 2015. minimac2: faster genotype imputation. Bioinformatics 31:782–784. https://doi.org/10.1093/bioinformatics/btu704.

Gamazon, E. R., H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E. Eyler, J. C. Denny, D. L. Nicolae, N. J. Cox, H. K. Im, and GTEx Consortium. 2015. A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. 47:1091–1098. https://doi.org/10.1038/ng.3367.

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics, selection, evolution. Genet. Sel. Evol. 41:55.

Gautason, E., A. A. Schonherz, G. Sahana, and B. Guldbrandtsen. 2021. Genomic inbreeding and selection signatures in the local dairy breed Icelandic Cattle. Anim. Genet. 52:251–262. https://doi.org/10.1111/age.13058.

Gibson, G. 2012. Rare and common variants: twenty arguments. Nat. Rev. Genet. 13:135–145. https://doi.org/10.1038/nrg3118.

Gilad, Y., S. A. Rifkin, and J. K. Pritchard. 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 24:408–415. https://doi.org/10.1016/j.tig.2008.06.001.

Goeman, J. J., and A. Solari. 2014. Multiple hypothesis testing in genomics. Stat. Med. 33:1946–1978. https://doi.org/10.1002/sim.6082.

Gogarten, S. M., T. Bhangale, M. P. Conomos, C. A. Laurie, C. P. McHugh, I. Painter, X. Zheng, D. R. Crosslin, D. Levine, T. Lumley, S. C. Nelson, K. Rice, J. Shen, R. Swarnkar, B. S. Weir, and C. C. Laurie. 2012. GWASTools: An R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics 28:3329–3331. https://doi.org/10.1093/bioinformatics/bts610.

González, J. R., L. Armengol, X. Solé, E. Guinó, J. M. Mercader, X. Estivill, and V. Moreno. 2007. SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23:654–655. https://doi.org/10.1093/bioinformatics/btm025.

Goring, H. H., J. D. Terwilliger, and J. Blangero. 2001. Large upward bias in estimation of locus-specific effects from genomewide scans. Am. J. Hum. Genet. 69:1357–1369. https://doi.org/10.1086/324471.

Gualdrón Duarte, J. L., R. J. C. Cantet, R. O. Bates, C. W. Ernst, N. E. Raney, and J. P. Steibel. 2014. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinformatics 15:246. https://doi.org/10.1186/1471-2105-15-246.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186. https://doi.org/10.1186/1471-2105-12-186.

Hayes, B., and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics, selection, evolution. Genet. Sel. Evol. 33:209–229.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443. https://doi.org/10.3168/jds.2008-1646.

Hendricks, A. E., J. Dupuis, M. W. Logue, R. H. Myers, and K. L. Lunetta. 2014. Correction for multiple testing in a gene region. Eur. J. Hum. Genet. 22:414–418. https://doi.org/10.1038/ejhg.2013.144.

Hormozdiari, F., E. Kostem, E. Y. Kang, B. Pasaniuc, and E. Eskin. 2014. Identifying causal variants at loci with multiple signals of association. Genetics 198:497–508. https://doi.org/10.1534/genetics.114.167908.

Hormozdiari, F., M. van de Bunt, A. V. Segré, X. Li, J. W. J. Joo, M. Bilow, J. H. Sul, S. Sankararaman, B. Pasaniuc, and E. Eskin. 2016. Colocalization of GWAS and eQTL signals detects target genes. Am. J. Hum. Genet. 99:1245–1260. https://doi.org/10.1016/j.ajhg.2016.10.003.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44:955–959. https://doi.org/10.1038/ng.2354.

Howie, B. N., P. Donnelly, and J. Marchini. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5:e1000529. https://doi.org/10.1371/journal.pgen.1000529.

Hu, Y., M. Li, Q. Lu, H. Weng, J. Wang, S. M. Zekavat, Z. Yu, B. Li, J. Gu, S. Muchnik, Y. Shi, B. W. Kunkle, S. Mukherjee, P. Natarajan, A. Naj, A. Kuzma, Y. Zhao, P. K. Crane, H. Lu, and H. Zhao. 2019. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat. Genet. 51:568–576. https://doi.org/10.1038/s41588-019-0345-7.

Hu, Z.-L., C. A. Park, and J. M. Reecy. 2022. Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. Nucleic Acids Res. 50(D1):D956–D961. https://doi.org/10.1093/nar/gkab1116.

Iotchkova, V., G. R. S. Ritchie, M. Geihs, S. Morganella, J. L. Min, K. Walter, N. J. Timpson, I. Dunham, E. Birney, and N. Soranzo. U. K. Consortium. 2019. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat. Genet. 51:343–353. https://doi.org/10.1038/s41588-018-0322-6.

Jiang, J., L. Ma, D. Prakapenka, P. M. VanRaden, J. B. Cole, and Y. Da. 2019a. A large-scale genome-wide association study in U.S. Holstein cattle. Front. Genet. 10:412. https://doi.org/10.3389/fgene.2019.00412.

Jiang, L., Z. Zheng, T. Qi, K. E. Kemper, N. R. Wray, P. M. Visscher, and J. Yang. 2019b. A resource-efficient tool for mixed model association analysis of large-scale data. Nat. Genet. 51:1749–1755. https://doi.org/10.1038/s41588-019-0530-8.

Kanehisa, M., and S. Goto. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27–30. https://doi.org/10.1093/nar/28.1.27.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42:348–354. https://doi.org/10.1038/ng.548.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723. https://doi.org/10.1534/genetics.107.080101.

Kemper, K. E., P. M. Visscher, and M. E. Goddard. 2012. Genetic architecture of body size in mammals. Genome Biol. 13:244. https://doi.org/10.1186/gb-2012-13-4-244.

Kern, C., Y. Wang, X. Xu, Z. Pan, M. Halstead, G. Chanthavixay, P. Saelao, S. Waters, R. Xiang, A. Chamberlain, I. Korf, M. E. Delany, H. H. Cheng, J. F. Medrano, A. L. Van Eenennaam, C. K. Tuggle, C. Ernst, P. Flicek, G. Quon, P. Ross, and H. Zhou. 2021. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nat. Commun. 12:1821. https://doi.org/10.1038/s41467-021-22100-8.

Khamis, A. M., O. Motwalli, R. Oliva, B. R. Jankovic, Y. A. Medvedeva, H. Ashoor, M. Essack, X. Gao, and V. B. Bajic. 2018. A novel method for improved accuracy of transcription factor binding site prediction. Nucleic Acids Res. 46:e72. https://doi.org/10.1093/nar/gky237.

Khatkar, M. S., G. Moser, B. J. Hayes, and H. W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13:538. https://doi.org/10.1186/1471-2164-13-538.

Kichaev, G., W. Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L. Price, P. Kraft, and B. Pasaniuc. 2014. Integrating functional

data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 10:e1004722. https://doi.org/10.1371/journal.pgen.1004722.

Knüppel, R., P. Dietze, W. Lehnberg, K. Frech, and E. Wingender. 1994. TRANSFAC retrieval program: A network model database of eukaryotic transcription regulating sequences and proteins. J. Comput. Biol. 1:191–198. https://doi.org/10.1089/cmb.1994.1.191.

Korkuć, P., D. Arends, and G. A. Brockmann. 2019. Finding the optimal imputation strategy for small cattle populations. Front. Genet. 10:52. https://doi.org/10.3389/fgene.2019.00052.

Lamparter, D., D. Marbach, R. Rueedi, Z. Kutalik, and S. Bergmann. 2016. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLOS Comput. Biol. 12:e1004714. https://doi.org/10.1371/journal.pcbi.1004714.

Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756. https://doi.org/10.1093/genetics/124.3.743.

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, T. Ferreira, A. R. Wood, R. J. Weyant, A. V. Segrè, E. K. Speliotes, E. Wheeler, N. Soranzo, J.-H. Park, J. Yang, D. Gudbjartsson, N. L. Heard-Costa, J. C. Randall, L. Qi, A. Vernon Smith, R. Mägi, T. Pastinen, L. Liang, I. M. Heid, J. A. Luan, G. Thorleifsson, T. W. Winkler, M. E. Goddard, K. Sin Lo, C. Palmer, T. Workalemahu, Y. S. Aulchenko, Å. Johansson, M. Carola Zillikens, M. F. Feitosa, T. Esko, T. Johnson, S. Ketkar, P. Kraft, M. Mangino, I. Prokopenko, D. Absher, E. Albrecht, F. Ernst, N. L. Glazer, C. Hayward, J.-J. Hottenga, K. B. Jacobs, J. W. Knowles, Z. Kutalik, K. L. Monda, O. Polasek, M. Preuss, N. W. Rayner, N. R. Robertson, V. Steinthorsdottir, J. P. Tyrer, B. F. Voight, F. Wiklund, J. Xu, J. Hua Zhao, D. R. Nyholt, N. Pellikka, M. Perola, J. R. B. Perry, I. Surakka, M.-L. Tammesoo, E. L. Altmaier, N. Amin, T. Aspelund, T. Bhangale, G. Boucher, D. I. Chasman, C. Chen, L. Coin, M. N. Cooper, A. L. Dixon, Q. Gibson, E. Grundberg, K. Hao, M. Juhani Junttila, L. M. Kaplan, J. Kettunen, I. R. König, T. Kwan, R. W. Lawrence, D. F. Levinson, M. Lorentzon, B. McKnight, A. P. Morris, M. Müller, J. Suh Ngwa, S. Purcell, S. Rafelt, R. M. Salem, E. Salvi, S. Sanna, J. Shi, U. Sovio, J. R. Thompson, M. C. Turchin, L. Vandenput, D. J. Verlaan, V. Vitart, C. C. White, A. Ziegler, P. Almgren, A. J. Balmforth, H. Campbell, L. Citterio, A. De Grandi, A. Dominiczak, J. Duan, P. Elliott, R. Elosua, J. G. Eriksson, N. B. Freimer, E. J. C. Geus, N. Glorioso, S. Haiqing, A.-L. Hartikainen, A. S. Havulinna, A. A. Hicks, J. Hui, W. Igl, T. Illig, A. Jula, E. Kajantie, T. O. Kilpeläinen, M. Koiranen, I. Kolcic, S. Koskinen, P. Kovacs, J. Laitinen, J. Liu, M.-L. Lokki, A. Marusic, A. Maschio, T. Meitinger, A. Mula, G. Paré, A. N. Parker, J. F. Peden, A. Petersmann, I. Pichler, K. H. Pietiläinen, A. Pouta, M. Ridderstråle, J. I. Rotter, J. G. Sambrook, A. R. Sanders, C. Oliver Schmidt, J. Sinisalo, J. H. Smit, H. M. Stringham, G. Bragi Walters, E. Widen, S. H. Wild, G. Willemsen, L. Zagato, L. Zgaga, P. Zitting, H. Alavere, M. Farrall, W. L. McArdle, M. Nelis, M. J. Peters, S. Ripatti, J. B. J. van Meurs, K. K. Aben, K. G. Ardlie, J. S. Beckmann, J. P. Beilby, R. N. Bergman, S. Bergmann, F. S. Collins, D. Cusi, M. den Heijer, G. Eiriksdottir, P. V. Gejman, A. S. Hall, A. Hamsten, H. V. Huikuri, C. Iribarren, M. Kähönen, J. Kaprio, S. Kathiresan, L. Kiemeney, T. Kocher, L. J. Launer, T. Lehtimäki, O. Melander, T. H. Mosley Jr., A. W. Musk, M. S. Nieminen, C. J. O'Donnell, C. Ohlsson, B. Oostra, L. J. Palmer, O. Raitakari, P. M. Ridker, J. D. Rioux, A. Rissanen, C. Rivolta, H. Schunkert, A. R. Shuldiner, D. S. Siscovick, M. Stumvoll, A. Tönjes, J. Tuomilehto, G.-J. van Ommen, J. Viikari, A. C. Heath, N. G. Martin, G. W. Montgomery, M. A. Province, M. Kayser, A. M. Arnold, L. D. Atwood, E. Boerwinkle, S. J. Chanock, P. Deloukas, C. Gieger, H. Grönberg, P. Hall, A. T. Hattersley, C. Hengstenberg, W. Hoffman, G. Mark Lathrop, V. Salomaa, S. Schreiber, M. Uda, D. Waterworth, A. F. Wright, T. L. Assimes, I. Barroso, A. Hofman, K. L. Mohlke, D. I. Boomsma, M. J. Caulfield, L. Adrienne Cupples, J. Erdmann, C. S. Fox, V. Gudnason, U. Gyllensten, T. B. Harris, R. B. Hayes, M.-R. Jarvelin, V. Mooser, P. B. Munroe, W. H. Ouwehand, B. W. Penninx, P. P. Pramstaller, T. Quertermous, I. Rudan, N. J. Samani, T. D. Spector, H. Völzke, H. Watkins, J. F. Wilson, L. C. Groop, T. Haritunians, F. B. Hu, R. C. Kaplan, A. Metspalu, K. E. North, D. Schlessinger, N. J. Wareham, D. J. Hunter, J. R. O'Connell, D. P. Strachan, H. E. Wichmann, I. B. Borecki, C. M. van Duijn, E. E. Schadt, U. Thorsteinsdottir, L. Peltonen, A. G. Uitterlinden, P. M. Visscher, N. Chatterjee, R. J. F. Loos, M. Boehnke, M. I. McCarthy, E. Ingelsson, C. M. Lindgren, G. R. Abecasis, K. Stefansson, T. M. Frayling, and J. N. Hirschhorn. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832–838.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656–4663. https://doi.org/10.3168/jds.2009-2061.

Legarra, A., and V. Ducrocq. 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. J. Dairy Sci. 95:4629–4645. https://doi.org/10.3168/jds.2011-4982.

Li, J., and L. Ji. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity 95:221–227. https://doi.org/10.1038/sj.hdy.6800717.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–293. https://doi.org/10.1126/science.1181369.

Listgarten, J., C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin, and D. Heckerman. 2012. Improved linear mixed models for genome-wide association studies. Nat. Methods 9:525–526. https://doi.org/10.1038/nmeth.2037.

Liu, J. Z., A. F. McRae, D. R. Nyholt, S. E. Medland, N. R. Wray, K. M. Brown, N. K. Hayward, G. W. Montgomery, P. M. Visscher, N. G. Martin, and S. Macgregor. 2010. A versatile gene-based test for genome-wide association studies. Am. J. Hum. Genet. 87:139–145. https://doi.org/10.1016/j.ajhg.2010.06.009.

Liu, S., Y. Gao, O. Canela-Xandri, S. Wang, Y. Yu, W. Cai, B. Li, R. Xiang, A. J. Chamberlain, E. Pairo-Castineira, K. D'Mellow, K. Rawlik, C. Xia, Y. Yao, P. Navarro, D. Rocha, X. Li, Z. Yan, C. Li, B. D. Rosen, C. P. Van Tassell, P. M. Vanraden, S. Zhang, L. Ma, J. B. Cole, G. E. Liu, A. Tenesa, and L. Fang. 2022. A multi-tissue atlas of regulatory variants in cattle. Nat. Genet. 54:1438–1447. https://doi.org/10.1038/s41588-022-01153-5.

Loh, P.-R., P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A. Reshef, H. K. Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, R. Durbin, and A. L. Price. 2016a. Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. 48:1443–1448. https://doi.org/10.1038/ng.3679.

Loh, P.-R., P. F. Palamara, and A. L. Price. 2016b. Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. 48:811–816. https://doi.org/10.1038/ng.3571.

Lund, M. S., A. P. Roos, A. G. Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, F. Seefried, and G. Su. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet. Sel. Evol. 43:43.

MacArthur, D. G., T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. A. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. A. Pennacchio, J. A. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler, and C. Gunter. 2014. Guidelines for investigating causality of sequence variants in human disease. Nature 508:469–476. https://doi.org/10.1038/nature13127.

MacArthur, J., E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, and H. Parkinson. 2017. The new NHGRI-EBI Catalog of published

genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45(D1):D896–D901. https://doi.org/10.1093/nar/gkw1133.
Madsen, B. E., and S. R. Browning. 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 5:e1000384. https://doi.org/10.1371/journal.pgen.1000384.
Mägi, R., M. Horikoshi, T. Sofer, A. Mahajan, H. Kitajima, N. Franceschini, M. I. McCarthy, and A. P. Morris. 2017. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. Hum. Mol. Genet. 26:3639–3650. https://doi.org/10.1093/hmg/ddx280.
Makanjuola, B. O., F. Miglior, E. A. Abdalla, C. Maltecca, F. S. Schenkel, and C. F. Baes. 2020. Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations. J. Dairy Sci. 103:5183–5199. https://doi.org/10.3168/jds.2019-18013.
Mao, X., G. Sahana, D. J. De Koning, and B. Guldbrandtsen. 2016. Genome-wide association studies of growth traits in three dairy cattle breeds using whole-genome sequence data. J. Anim. Sci. 94:1426–1437. https://doi.org/10.2527/jas.2015-9838.
Mao, X., G. Sahana, A. M. Johansson, A. Liu, A. Ismael, P. Lovendahl, D. J. De Koning, and B. Guldbrandtsen. 2020. Genome-wide association mapping for dominance effects in female fertility using real and simulated data from Danish Holstein cattle. Sci. Rep. 10:2953. https://doi.org/10.1038/s41598-020-59788-5.
Marchini, J., L. R. Cardon, M. S. Phillips, and P. Donnelly. 2004. The effects of human population structure on large genetic association studies. Nat. Genet. 36:512–517. https://doi.org/10.1038/ng1337.
Marchini, J., and B. Howie. 2010. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11:499–511. https://doi.org/10.1038/nrg2796.
Marees, A. T., H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, and E. M. Derks. 2018. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int. J. Methods Psychiatr. Res. 27:e1608. https://doi.org/10.1002/mpr.1608.
Mattick, J. S., and I. V. Makunin. 2006. Non-coding RNA. Hum. Mol. Genet. 15(Suppl. 1):R17–R29. https://doi.org/10.1093/hmg/ddl046.
McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17:122. https://doi.org/10.1186/s13059-016-0974-4.
Mesbah-Uddin, M., B. Guldbrandtsen, A. Capitan, M. S. Lund, D. Boichard, and G. Sahana. 2022. Genome-wide association study with imputed whole-genome sequence variants including large deletions for female fertility in 3 Nordic dairy cattle breeds. J. Dairy Sci. 105:1298–1313. https://doi.org/10.3168/jds.2021-20655.
Mesbah-Uddin, M., C. Hoze, P. Michot, A. Barbat, R. Lefebvre, M. Boussaha, G. Sahana, S. Fritz, D. Boichard, and A. Capitan. 2019. A missense mutation (p.Tyr452Cys) in the CAD gene compromises reproductive success in French Normande cattle. J. Dairy Sci. 102:6340–6356. https://doi.org/10.3168/jds.2018-16100.
Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819.
Michot, P., S. Chahory, A. Marete, C. Grohs, D. Dagios, E. Donzel, A. Aboukadiri, M. C. Deloche, A. Allais-Bonnet, M. Chambrial, S. Barbey, L. Genestout, M. Boussaha, C. Danchin-Burge, S. Fritz, D. Boichard, and A. Capitan. 2016. A reverse genetic approach identifies an ancestral frameshift mutation in RP1 causing recessive progressive retinal degeneration in European cattle breeds. Genet. Sel. Evol. 48:56. https://doi.org/10.1186/s12711-016-0232-y.
Mishra, A., and S. Macgregor. 2015. VEGAS2: Software for more flexible gene-based testing. Twin Res. Hum. Genet. 18:86–91. https://doi.org/10.1017/thg.2014.79.
Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. J. Dairy Sci. 97:3943–3952. https://doi.org/10.3168/jds.2013-7752.
Mulder, H. A., M. P. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95:876–889. https://doi.org/10.3168/jds.2011-4490.
Muranty, H., and B. Goffinet. 1997. Selective genotyping for location and estimation of the effect of a quantitative trait locus. Biometrics 53:629–643. https://doi.org/10.2307/2533963.
Ng, P. C., and S. Henikoff. 2002. Accounting for human polymorphisms predicted to affect protein function. Genome Res. 12:436–446. https://doi.org/10.1101/gr.212802.
Nicolazzi, E. L., M. Picciolini, F. Strozzi, R. D. Schnabel, C. Lawley, A. Pirani, F. Brew, and A. Stella. 2014. SNPchiMp: A database to disentangle the SNPchip jungle in bovine livestock. BMC Genomics 15:123. https://doi.org/10.1186/1471-2164-15-123.
O'Connell, J., D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M. Traglia, J. Huang, J. E. Huffman, I. Rudan, R. McQuillan, R. M. Fraser, H. Campbell, O. Polasek, G. Asiki, K. Ekoru, C. Hayward, A. F. Wright, V. Vitart, P. Navarro, J. F. Zagury, J. F. Wilson, D. Toniolo, P. Gasparini, N. Soranzo, M. S. Sandhu, and J. Marchini. 2014. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 10:e1004234. https://doi.org/10.1371/journal.pgen.1004234.
Palmer, C., and I. Pe'er. 2017. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. PLoS Genet. 13:e1006916. https://doi.org/10.1371/journal.pgen.1006916.
Palmer, D. S., W. Zhou, L. Abbott, N. Baya, C. Churchhouse, C. Seed, T. Poterba, D. King, M. Kanai, A. Bloemendal, and B. M. Neale. 2022. Analysis of genetic dominance in the UK Biobank. bioRxiv: 2021.2008.2015.456387.
Panagiotou, O. A., and J. P. A. Ioannidis. 2012. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int. J. Epidemiol. 41:273–286. https://doi.org/10.1093/ije/dyr178.
Pe'er, I., R. Yelensky, D. Altshuler, and M. J. Daly. 2008. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet. Epidemiol. 32:381–385. https://doi.org/10.1002/gepi.20303.
Pook, T., M. Mayer, J. Geibel, S. Weigend, D. Cavero, C. C. Schoen, and H. Simianer. 2020. Improving imputation quality in BEAGLE for crop and livestock data. G3 (Bethesda) 10:177–188. https://doi.org/10.1534/g3.119.400798.
Price, A. L., G. V. Kryukov, P. I. W. de Bakker, S. M. Purcell, J. Staples, L. J. Wei, and S. R. Sunyaev. 2010a. Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. 86:832–838. https://doi.org/10.1016/j.ajhg.2010.04.005.
Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909. https://doi.org/10.1038/ng1847.
Price, A. L., N. A. Zaitlen, D. Reich, and N. Patterson. 2010b. New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11:459–463. https://doi.org/10.1038/nrg2813.
Pruim, R. J., R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines, T. P. Gliedt, M. Boehnke, G. R. Abecasis, and C. J. Willer. 2010. LocusZoom: Regional visualization of genome-wide association scan results. Bioinformatics 26:2336–2337. https://doi.org/10.1093/bioinformatics/btq419.
Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575. https://doi.org/10.1086/519795.
Qanbari, S., R. D. Schnabel, and D. Wittenburg. 2022. Evidence of rare misassemblies in the bovine reference genome revealed by population genetic metrics. Anim. Genet. 53:498–505. https://doi.org/10.1111/age.13205.
Qi, T., Y. Wu, J. Zeng, F. Zhang, A. Xue, L. Jiang, Z. Zhu, K. Kemper, L. Yengo, and Z. Zheng. eQTLGen Consortium, R. E. Marioni, G. W. Montgomery, I. J. Deary, N. R. Wray, P. M. Visscher, A.

F. McRae, and J. Yang. 2018. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. Nat. Commun. 9:2282. https://doi.org/10.1038/s41467-018-04558-1.

Reed, E., S. Nunez, D. Kulp, J. Qian, M. P. Reilly, and A. S. Foulkes. 2015. A guide to genome-wide association analysis and post-analytic interrogation. Stat. Med. 34:3769–3792. https://doi.org/10.1002/sim.6605.

Reis-Filho, J. S. 2009. Next-generation sequencing. Breast Cancer Res. 11(Suppl. 3):S12.

Reynolds, E. G. M., T. Lopdell, Y. Wang, K. M. Tiplady, C. S. Harland, T. J. J. Johnson, C. Neeley, K. Carnie, R. G. Sherlock, C. Couldrey, S. R. Davis, B. L. Harris, R. J. Spelman, D. J. Garrick, and M. D. Littlejohn. 2022. Non-additive QTL mapping of lactation traits in 124,000 cattle reveals novel recessive loci. Genet. Sel. Evol. 54:5. https://doi.org/10.1186/s12711-021-00694-3.

Reynolds, E. G. M., C. Neeley, T. J. Lopdell, M. Keehan, K. Dittmer, C. S. Harland, C. Couldrey, T. J. J. Johnson, K. Tiplady, G. Worth, M. Walker, S. R. Davis, R. G. Sherlock, K. Carnie, B. L. Harris, C. Charlier, M. Georges, R. J. Spelman, D. J. Garrick, and M. D. Littlejohn. 2021. Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. Nat. Genet. 53:949–954. https://doi.org/10.1038/s41588-021-00872-5.

Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517. https://doi.org/10.1126/science.273.5281.1516.

Robinson, J. T., H. Thorvaldsdóttir, A. M. Wenger, A. Zehir, and J. P. Mesirov. 2017. Variant review with the integrative genomics viewer. Cancer Res. 77:e31–e34. https://doi.org/10.1158/0008-5472.CAN-17-0337.

Rosen, B. D., D. M. Bickhart, R. D. Schnabel, S. Koren, C. G. Elsik, E. Tseng, T. N. Rowan, W. Y. Low, A. Zimin, C. Couldrey, R. Hall, W. Li, A. Rhie, J. Ghurye, S. D. McKay, F. Thibaud-Nissen, J. Hoffman, B. M. Murdoch, W. M. Snelling, T. G. McDaneld, J. A. Hammond, J. C. Schwartz, W. Nandolo, D. E. Hagen, C. Dreischer, S. J. Schultheiss, S. G. Schroeder, A. M. Phillippy, J. B. Cole, C. P. Van Tassell, G. Liu, T. P. L. Smith, and J. F. Medrano. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience 9:giaa021. https://doi.org/10.1093/gigascience/giaa021.

Rotroff, D. M. 2020. A bioinformatics crash course for interpreting genomics data. Chest 158(Suppl. 1):S113–S123. https://doi.org/10.1016/j.chest.2020.03.004.

Sahana, G., B. Guldbrandtsen, L. Janss, and M. S. Lund. 2010. Comparison of association mapping methods in a complex pedigreed population. Genet. Epidemiol. 34:455–462. https://doi.org/10.1002/gepi.20499.

Salyakina, D., S. R. Seaman, B. L. Browning, F. Dudbridge, and B. Muller-Myhsok. 2005. Evaluation of Nyholt's procedure for multiple testing correction. Hum. Hered. 60:19–25. https://doi.org/10.1159/000087540.

Sanchez, M. P., Y. Ramayo-Caldas, V. Wolf, C. Laithier, M. El Jabri, A. Michenet, M. Boussaha, S. Taussat, S. Fritz, A. Delacroix-Buchet, M. Brochard, and D. Boichard. 2019. Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbeliarde cows. Genet. Sel. Evol. 51:34. https://doi.org/10.1186/s12711-019-0473-7.

Sanchez, M.-P., D. Rocha, M. Charles, M. Boussaha, C. Hozé, M. Brochard, A. Delacroix-Buchet, P. Grosperrin, and D. Boichard. 2021. Sequence-based GWAS and post-GWAS analyses reveal a key role of SLC37A1, ANKH, and regulatory regions on bovine milk mineral content. Sci. Rep. 11:7537. https://doi.org/10.1038/s41598-021-87078-1.

Sanchez, M.-P., T. Tribout, S. Fritz, R. Guatteo, C. Fourichon, L. Schibler, A. Delafosse, and D. Boichard. 2022. New insights into the genetic resistance to paratuberculosis in Holstein cattle via single-step genomic evaluation. Genet. Sel. Evol. 54:67. https://doi.org/10.1186/s12711-022-00757-z.

Sandelin, A., W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard. 2004. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 32:D91–D94. https://doi.org/10.1093/nar/gkh012.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478. https://doi.org/10.1186/1471-2164-15-478.

Sawicki, M. P., G. Samara, M. Hurwitz, and E. Passaro Jr.. 1993. Human Genome Project. Am. J. Surg. 165:258–264. https://doi.org/10.1016/S0002-9610(05)80522-7.

Schaid, D. J., W. Chen, and N. B. Larson. 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet. 19:491–504. https://doi.org/10.1038/s41576-018-0016-z.

Schulz, K. F., and D. A. Grimes. 2002. Case-control studies: Research in reverse. Lancet 359:431–434. https://doi.org/10.1016/S0140-6736(02)07605-5.

Sesia, M., S. Bates, E. Candes, J. Marchini, and C. Sabatti. 2021. False discovery rate control in genome-wide association studies with population structure. Proc. Natl. Acad. Sci. USA 118:e2105841118. https://doi.org/10.1073/pnas.2105841118.

Sham, P. C., and S. M. Purcell. 2014. Statistical power and significance testing in large-scale genetic studies. Nat. Rev. Genet. 15:335–346. https://doi.org/10.1038/nrg3706.

Speed, D., J. Holmes, and D. J. Balding. 2020. Evaluating and improving heritability models using summary statistics. Nat. Genet. 52:458–462. https://doi.org/10.1038/s41588-020-0600-y.

Spencer, C. C., Z. Su, P. Donnelly, and J. Marchini. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 5:e1000477. https://doi.org/10.1371/journal.pgen.1000477.

Stephens, M., and D. J. Balding. 2009. Bayesian statistical methods for genetic association studies. Nat. Rev. Genet. 10:681–690. https://doi.org/10.1038/nrg2615.

Storey, J. D. 2002. A direct approach to false discovery rates. J. R. Stat. Soc. Series B Stat. Methodol. 64:479–498. https://doi.org/10.1111/1467-9868.00346.

Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA 100:9440–9445. https://doi.org/10.1073/pnas.1530509100.

Strandén, I., and D. J. Garrick. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92:2971–2975. https://doi.org/10.3168/jds.2008-1929.

Su, G., R. F. Brondum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (approximately 54,000) and high-density (approximately 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J. Dairy Sci. 95:4657–4665. https://doi.org/10.3168/jds.2012-5379.

Tarailo-Graovac, M., and N. Chen. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinform. 25:4.10.1–4.10.14.

Teo, Y. Y. 2008. Common statistical issues in genome-wide association studies: A review on power, data quality control, genotype calling and population structure. Curr. Opin. Lipidol. 19:133–143. https://doi.org/10.1097/MOL.0b013e3282f5dd77.

Tian, D., P. Wang, B. Tang, X. Teng, C. Li, X. Liu, D. Zou, S. Song, and Z. Zhang. 2020. GWAS Atlas: A curated resource of genome-wide variant-trait associations in plants and animals. Nucleic Acids Res. 48(D1):D927–D932. https://doi.org/10.1093/nar/gkz828.

Tiezzi, F., K. L. Parker-Gaddis, J. B. Cole, J. S. Clay, and C. Maltecca. 2015. A genome-wide association study for clinical mastitis in first parity US Holstein cows using single-step approach and genomic matrix re-weighting procedure. PLoS One 10:e0114919. https://doi.org/10.1371/journal.pone.0114919.

Tribout, T., P. Croiseau, R. Lefebvre, A. Barbat, M. Boussaha, S. Fritz, D. Boichard, C. Hoze, and M.-P. Sanchez. 2020. Confirmed effects of candidate variants for milk production, udder health, and udder morphology in dairy cattle. Genet. Sel. Evol. 52:55. https://doi.org/10.1186/s12711-020-00575-1.

Turley, P., R. K. Walters, O. Maghzian, A. Okbay, J. J. Lee, M. A. Fontana, T. A. Nguyen-Viet, R. Wedow, M. Zacher, N. A. Furlotte, P. Magnusson, S. Oskarsson, M. Johannesson, P. M. Visscher, D. Laibson, D. Cesarini, B. M. Neale, and D. J. Benjamin. 2018. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat. Genet. 50:229–237. https://doi.org/10.1038/s41588-017-0009-4.

Turner, S., L. L. Armstrong, Y. Bradford, C. S. Carlson, D. C. Crawford, A. T. Crenshaw, M. de Andrade, K. F. Doheny, J. L. Haines, G. Hayes, G. Jarvik, L. Jiang, I. J. Kullo, R. Li, H. Ling, T. A. Manolio, M. Matsumoto, C. A. McCarty, A. N. McDavid, D. B. Mirel, J. E. Paschall, E. W. Pugh, L. V. Rasmussen, R. A. Wilke, R. L. Zuvich, and M. D. Ritchie. 2011. Quality control procedures for genome-wide association studies. Curr. Protoc. Hum. Genet. 68:1.19.1–1.19.18.

Uffelmann, E., Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma. 2021. Genome-wide association studies. Nat. Rev. Methods Primers 1:59. https://doi.org/10.1038/s43586-021-00056-9.

van Binsbergen, R., M. C. A. M. Bink, M. P. L. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 46:41. https://doi.org/10.1186/1297-9686-46-41.

van den Berg, I., R. Xiang, J. Jenko, H. Pausch, M. Boussaha, C. Schrooten, T. Tribout, A. B. Gjuvsland, D. Boichard, Ø. Nordbø, M.-P. Sanchez, and M. E. Goddard. 2020. Meta-analysis for milk fat and protein percentage using imputed sequence variant genotypes in 94,321 cattle from eight cattle breeds. Genet. Sel. Evol. 52:37. https://doi.org/10.1186/s12711-020-00556-4.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo. 2013. From FastQ Data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr. Protoc. Hum. Genet. 43:11.10.11–11.10.33.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

VanRaden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal-model information. J. Dairy Sci. 74:2737–2746. https://doi.org/10.3168/jds.S0022-0302(91)78453-1.

Voight, B. F., and J. K. Pritchard. 2005. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 1:e32. https://doi.org/10.1371/journal.pgen.0010032.

Wang, G., A. Sarkar, P. Carbonetto, and M. Stephens. 2020. A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Series B Stat. Methodol. 82:1273–1300. https://doi.org/10.1111/rssb.12388.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. (Camb.) 94:73–83. https://doi.org/10.1017/S0016672312000274.

Wang, K., M. Li, and H. Hakonarson. 2010. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38:e164. https://doi.org/10.1093/nar/gkq603.

Wang, M. H., H. J. Cordell, and K. Van Steen. 2019. Statistical methods for genome-wide association studies. Semin. Cancer Biol. 55:53–60. https://doi.org/10.1016/j.semcancer.2018.04.008.

Weale, M. E. 2010. Quality control for genome-wide association studies. Pages 341–372 in Genetic Variation: Methods and Protocols. M. R. Barnes and G. Breen, ed. Humana Press.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678. https://doi.org/10.1038/nature05911.

Whalen, A., R. Ros-Freixedes, D. L. Wilson, G. Gorjanc, and J. M. Hickey. 2018. Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. Genet. Sel. Evol. 50:67. https://doi.org/10.1186/s12711-018-0438-2.

Willer, C. J., Y. Li, and G. R. Abecasis. 2010. METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26:2190–2191. https://doi.org/10.1093/bioinformatics/btq340.

Wu, X., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2016. Association analysis for feet and legs disorders with whole-genome sequence variants in 3 dairy cattle breeds. J. Dairy Sci. 99:7221–7231. https://doi.org/10.3168/jds.2015-10705.

Xu, S. 2003. Theoretical basis of the Beavis effect. Genetics 165:2259–2268. https://doi.org/10.1093/genetics/165.4.2259.

Yang, J., T. Ferreira, A. P. Morris, S. E. Medland, P. A. F. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, R. J. Loos, T. M. Frayling, M. I. McCarthy, J. N. Hirschhorn, M. E. Goddard, and P. M. Visscher. 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. 44:369–375. https://doi.org/10.1038/ng.2213.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011a. GCTA: A tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

Yang, J., M. N. Weedon, S. Purcell, G. Lettre, K. Estrada, C. J. Willer, A. V. Smith, E. Ingelsson, J. R. O'Connell, M. Mangino, R. Magi, P. A. Madden, A. C. Heath, D. R. Nyholt, N. G. Martin, G. W. Montgomery, T. M. Frayling, J. N. Hirschhorn, M. I. McCarthy, M. E. Goddard, and P. M. Visscher.GIANT Consortium. 2011b. Genomic inflation factors under polygenic inheritance. Eur. J. Hum. Genet. 19:807–812. https://doi.org/10.1038/ejhg.2011.39.

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price. 2014. Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. 46:100–106. https://doi.org/10.1038/ng.2876.

Yin, T., and S. König. 2019. Genome-wide associations and detection of potential candidate genes for direct genetic and maternal genetic effects influencing dairy cattle body weight at different ages. Genet. Sel. Evol. 51:4. https://doi.org/10.1186/s12711-018-0444-4.

Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208. https://doi.org/10.1038/ng1702.

Zhang, Q., B. Guldbrandtsen, M. P. Calus, M. S. Lund, and G. Sahana. 2016. Comparison of gene-based rare variant association mapping methods for quantitative traits in a bovine population with complex familial relationships. Genet. Sel. Evol. 48:60. https://doi.org/10.1186/s12711-016-0238-5.

Zhao, H., N. Mitra, P. A. Kanetsky, K. L. Nathanson, and T. R. Rebbeck. 2018. A practical approach to adjusting for population stratification in genome-wide association studies: principal components and propensity scores (PCAPS). Stat. Appl. Genet. Mol. Biol. 17:6. https://doi.org/10.1515/sagmb-2017-0054.

Zhu, Z., F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher, and J. Yang. 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. 48:481–487. https://doi.org/10.1038/ng.3538.

Zondervan, K. T., and L. R. Cardon. 2007. Designing candidate gene and genome-wide case-control association studies. Nat. Protoc. 2:2492–2501. https://doi.org/10.1038/nprot.2007.366.

ORCIDS

G. Sahana https://orcid.org/0000-0001-7608-7577
Z. Cai https://orcid.org/0000-0002-9579-3415
M. P. Sanchez https://orcid.org/0000-0002-1371-5342
A. C. Bouwman https://orcid.org/0000-0001-5079-7108
D. Boichard https://orcid.org/0000-0003-0361-2961

APPENDIX 1

GBLUP and SNP-BLUP Approaches for GWAS

The GBLUP and SNP-BLUP approaches estimate all SNP effects simultaneously for all genotyped individuals with performances. Although SNP effect estimates are strongly regressed, SNP-BLUP-derived test statistics for marker effects are equivalent to those obtained from single-SNP linear regression (e.g., by using EMMAX; Gualdrón Duarte et al., 2014; Chen et al., 2017). Because a QTL effect is distributed over several neighboring SNPs, SNP signals must be collected in intervals. As the GBLUP model is equivalent to SNP-BLUP, it can be used for the same purpose, after back-solving for SNP effects. This approach can also be extended to a population of genotyped and ungenotyped animals in the so-called single-step approach, thus allowing the analysis of complex data sets (Aguilar et al., 2019). Again, single-step GBLUP (Legarra et al., 2009; Christensen and Lund, 2010) and single-step SNP-BLUP (Legarra and Ducrocq, 2012; Wang et al., 2012; Fernando et al., 2014; Aguilar et al., 2019) can be used similarly, as their models are equivalent to estimate both breeding values and marker effects. Both approaches have been used for GWAS (Tiezzi et al., 2015; Aguilar et al., 2019).

However, GBLUP- and SNP-BLUP-based approaches have 2 important limitations:

(a) The SNP effects are very strongly regressed and the estimated effect of the top variant is much lower than the true effect. Indeed, the equation corresponding to SNP i is of the form $\hat{s} = \frac{\mathbf{m_i}' \mathbf{y_c}}{\mathbf{m_i}' \mathbf{m_i} + \lambda}$, where $\mathbf{m_i}$ is the vector of centered genotypes for SNP i, $\mathbf{y_c}$ is the vector of phenotypes adjusted for the effects of all other SNPs and other effects in the model, and $\lambda = \sigma_e^2 / \sigma_s^2 = 2\sum_j p_j(1-p_j)\frac{\sigma_e^2}{\sigma_g^2}$, with σ_e^2, σ_s^2, and σ_g^2 the residual, SNP, and genetic variances, respectively, and p_j the minor allele frequency of variant j. In many analyses, this coefficient λ is larger than the number of data and the SNP effect is strongly regressed toward zero. The QTL effect is thus distributed over many markers, in the QTL region but sometimes also at long distance. To obtain a realistic estimate of the variance associated with a QTL, it is necessary to combine the effects of all markers of the region, and even so, the explained variance is still underestimated (e.g., see Sanchez et al., 2022, for an illustration).

(b) GBLUP and SNP-BLUP are suitable methods for moderate numbers of markers, of the order of a few tens of thousands of markers. However, when GWAS analyses aim to identify the best candidate causal variants, this implies performing the analyses at the sequence level. With this very large number of WGS variants, the limitations described above are greatly amplified, preventing use of GBLUP in practice for association studies.

APPENDIX 2

Bayesian Methods Applied to GWAS

Bayesian methods provide an alternative approach to assessing associations and mitigate the limitations of *P*-values, but come with the cost of additional modeling assumptions (Stephens and Balding, 2009). Allowing different prior distributions, the Bayesian methods provide a way to allocate different variances to SNPs and therefore to highlight the SNP with the strongest effect: their effect is less regressed. In the models with SNP selection, their probability of inclusion in each category is a direct mapping method. A whole alphabet of methods has been proposed, including methods enabling the selection of SNPs [e.g., BayesB (Meuwissen et al., 2001), BayesC (Habier et al., 2011), BayesR (Erbe et al., 2012)] and variable SNP variances [e.g., BayesA (Meuwissen et al., 2001), BayesB, or BayesR].

Sahana et al. (2010) compared Bayesian multiple regression analysis with LMM GWAS. Fernando and Garrick (2013) described Bayesian methods to combine large numbers of genotyped and nongenotyped animals for whole-genome analysis. In spite of their useful properties, Bayesian methods have limitations at the sequence level. Each QTL signal is scattered over multiple neighboring markers in strong LD, and, as with the SNP-BLUP approach, SNP signals must be collected in intervals (Sahana et al., 2010). Chen et al. (2017) reported advantages of a variable selection approach using LD-based windows. The second limitation is the computing cost, and Bayesian methods at a whole-sequence level are currently not computationally feasible on a routine basis.