Functionally prioritised whole-genome sequence variants improve the accuracy ofgenomic prediction for heat tolerance

Evans K. Cheruiyot (19548367@students.latrobe.edu.au)
School of Applied Systems BioUniversity, Bundoora, Victoria 3083, Australia https://orcid.org/0000-0002-6403-7967

Mekonnen Haile-Mariam (Mekonnen.HaileMariam@agriculture.vic.gov.au)
Agriculture Victoria Research, AgriBio, Centre for AgriBiosciences, Bundoora, Victoria 3083, Australia

Benjamin G. Cocks
School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3083, Australia

Iona M. MacLeod
Agriculture Victoria Research, AgriBio, Centre for AgriBiosciences, Bundoora, Victoria 3083, Australia

Raphael Mrode
Scotland's Rural College, Edinburgh, United Kingdom

Jennie E. Pryce
School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3083, Australia

Research Article

Keywords: GWAS, genomic prediction, dairy cattle, SNPs, whole-genome sequence

Posted Date: June 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-598177/v1

Functionally prioritised whole-genome sequence variants improve the accuracy of genomic prediction for heat tolerance

Evans K. Cheruiyot[1,2*], Mekonnen Haile-Mariam[2*], Benjamin G. Cocks[1,2], Iona M. MacLeod[2], Raphael Mrode[3,4], Jennie E. Pryce[1,2]

[1]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3083, Australia

[2]Agriculture Victoria Research, AgriBio, Centre for AgriBiosciences, Bundoora, Victoria 3083, Australia

[3]International Livestock Research Institute, Nairobi, Kenya

[4]Scotland's Rural College, Edinburgh, United Kingdom

*Corresponding author (s)

HM: Mekonnen.HaileMariam@agriculture.vic.gov.au

EKC: 19548367@students.latrobe.edu.au

Abstract

Background: Heat tolerance is a trait of economic importance in the context of warm climates and the effects of global warming on livestock, production, reproduction, health, and well-being. It is desirable to improve the prediction accuracy for heat tolerance to help accelerate the genetic gain for this trait. This study investigated the improvement in prediction accuracy for heat tolerance when selected sets of sequence variants from a large genome-wide association study (GWAS) were incorporated into a standard 50k SNP panel used by the industry.

Methods: Over 40,000 dairy cattle (Holsteins, Jersey, and crossbreds) with genotype and phenotype data were analysed. The phenotypes used to measure an individual's heat tolerance were defined as the rate of milk production decline (slope traits for the yield of milk, fat, and protein) with a rising temperature-humidity index. We used Holstein and Jersey cows to select sequence variants linked to heat tolerance based on GWAS. We then investigated the accuracy of prediction when sets of these pre-selected sequence variants were added to the 50k industry SNP array used routinely for genomic evaluations in Australia. We used a bull reference set to develop the genomic prediction equations and then validated them in an independent set of Holsteins, Jersey, and crossbred cows. The genomic prediction analyses were performed using BayesR and BayesRC methods.

Results: The accuracy of genomic prediction for heat tolerance improved by up to 7%, 5%, and 10% in Holsteins, Jersey, and crossbred cows, respectively, when sets of selected sequence markers from Holsteins (i.e., single-breed QTL discovery set) were added to the 50k industry SNP panel. Using pre-selected sequence variants identified based on a combined set of Holstein and Jersey cows in a multi-breed QTL discovery, a set of 6,132 to 6,422 SNPs generally improved accuracy, especially in the Jersey validation set. Combining Holstein and Jersey bulls (multi-breed) in the reference set improved prediction accuracy compared to using only Holstein bulls in the reference set.

Conclusion: Informative sequence markers can be prioritised to improve the genetic prediction of heat tolerance in different breeds, and these variants, in addition to providing biological insight, have direct application in the development of customized SNP arrays or can be utilised via imputation into current SNP sets.

Introduction

Heat tolerance is the ability of an animal to maintain production and reproduction levels under hot and humid conditions. With increasing global warming effects on animal production, the desire to breed for resilience to heat is growing worldwide, in part, to meet the demand of the increasing human population while coping with the challenges of hot and ever-changing production environments [1]. Dairy cows are often prone to heat stress due to the elevated metabolic heat of lactation. Temperature and humidity exceeding the threshold considered as comfortable for the dairy cows and other farm animals can compromise production (reduced milk, growth, etc.), reproduction (e.g., reduced conception rates), and welfare (elevated thirst and hunger), leading to substantial economic losses [2].

Considerable research has been conducted in many countries to assess heat tolerance and performance in farm livestock, including measuring changes in core body temperatures (e.g., rectal, vaginal, rumen, etc.) and thermal indices (e.g., temperature-humidity index (**THI**)) [3]. To study the effect of THI on milk production of dairy cows, [4] introduced a method whereby daily milk records are merged with temperature-humidity data to measure the rate of milk decline associated with changes in heat stress. This method has been widely adopted in many countries [5-7] due to the availability of extensive test-day milk records from dairy farms and climate data from weather stations.

In Australia, [7] used test-day milk records (milk, fat, and protein yield) and climate data collected from across Australia's dairying regions to evaluate heat tolerance in dairy cattle, which culminated in the release to the dairy industry (through DataGene Ltd; (https://datagene.com.au/) of the first genomic breeding values for this trait in 2017, with a reliability of 38%. While current prediction estimates are promising, even a smaller lift in reliability is economically important to the wider industry since the genetic improvement is linearly related to the selection intensity, accuracy of estimated breeding values (EBVs), genetic variation and is inversely proportional to the generation interval [8, 9]. The accuracy is the only component that is influenced by research in different ways to drive genetic improvement for a given trait whereas the other components (selection intensity, genetic variation, generation cycle) are largely controlled by breeding companies and farmers.

Besides increasing the size of the reference population, one way to boost the accuracy is to increase the density of markers used in genomic predictions. However, replacing the marker SNP panels with the full set of whole-genome sequence variants has, in most cases, yielded

limited, or no appreciable increase in the prediction accuracy for various traits in cattle [10], sheep [11], and avian species [12]. A promising alternative, in which a substantial increase in prediction accuracy that has been realized in previous reports, has been to augment standard industry SNP panels (e.g., 50k SNP array) with a small set of informative or causal mutations selected for a trait [11, 13-15]. To fully maximize predictions, this approach requires a careful selection of informative markers. Thanks to the 1000 Bull Genomes project (Hayes and Daetwyler, 2019), it is now possible to use this sequence database to impute genotypes to the whole-genome sequence. This may facilitate a more accurate selection of highly informative variants for genomic predictions, especially for complex traits such as heat tolerance. Specifically, having a large sample size and high-resolution genotypes can help identify many causal variants with medium- and small-sized effects.

In addition to the sample size, the composition of the population used for discovering informative variants can impact genomic predictions for a trait. Several studies [e.g., 16, 17, 18] have found improved precision of locating candidate causal variants when using multi-breed than the single-breed population in GWAS, especially for QTLs that segregate across breeds. This is partly due to shorter LD in multi-breed than within-breed analyses [18]. In a simulated study, [14] demonstrated that using variants close to the causal mutations can improve genomic predictions. With real data, [19] found increased accuracy of prediction for stature when using candidate variants discovered from meta-GWAS of 17 cattle populations. In sheep, [11] reported enhanced accuracy for various production traits when using pre-selected variants from the QTL discovery set that comprised multiple breed compositions. Besides these studies and several others that used single-breed sets to discover variants for traits, e.g., [15, 20], there is still a dearth of information on the value of variants from multi-breed populations in genomic predictions. Notably, it is critical to ensure that the population(s) used to discover informative sequence variants for a trait is independent of that used to train subsequent genomic predictions to avoid bias, as demonstrated by [21].

The main objective of this study was to investigate the prediction accuracy of heat tolerance in Holsteins when sets of selected sequence markers from a GWAS of a large sample size of Holstein cows were added to the 50k industry SNP panel used routinely for genomic evaluations in Australia. The selected variants are likely linked to causal mutations underpinning the genetic basis for heat tolerance [22] and, therefore, could enable more accurate and sustained genomic selection for heat tolerance. In addition, we investigated the accuracy of prediction when informative sequence markers discovered in Holstein cows are

used in the genomic predictions of numerically smaller breeds, including Jersey and crossbred cattle. Moreover, we investigated the gain in accuracy of prediction when using informative markers discovered in a combined set of Holstein and Jersey cows (i.e., multi-breed QTL discovery set). Finally, we compared the gain in accuracy when single-breed (Holstein bulls) versus multi-breed (Holstein + Jersey bulls) reference sets are used in the genomic predictions.

Materials and methods

Phenotypes

The phenotypes used were obtained from DataGene (DataGene Ltd., Melbourne, Australia; https://datagene.com.au/) and included test-day milk, fat, and protein yield for Holstein, Jersey, and Holstein-Jersey crossbred cows collected from dairy herds between 2003 and 2017 that were matched with climate data (daily temperature and humidity) obtained from weather stations across Australia's dairying regions. The distribution of dairy herds and weather stations, data filtering, and the calculation of environmental covariate (i.e., temperature-humidity index (**THI**)) used in this work was described in our earlier studies (Nguyen et al., 2016, Cheruiyot et al., 2020).

The rate of decline (slope) in milk, fat, and protein yield due to heat stress events was estimated using reaction norm models described by [23]. Briefly, data on milk, fat, or protein yield were adjusted for the fixed effects, including herd-test-day, year-season of calving, parity, Legendre polynomials (order 3) on the cow age on the day of the test, and the Legendre polynomials (order 8) on the interaction between parity and DIM. Random effects fitted in the model included a random regression on a linear orthogonal polynomial of THI, where the intercept represents the level of mean milk yield, and the linear component represents the change in milk yield (slope) due to heat stress for each cow and a residual term. The analyses to derive trait deviations (**TD**, which represents a phenotype adjusted for all fixed effects (i.e., the slope for each cow) were conducted using ASReml v4.2 (Gilmour et al., 2015). Slope solutions (i.e., TDs) for each bull's daughters were averaged to obtain heat tolerance slope traits for bulls and were equivalent to daughter trait deviations (**DTD**). From here on, the slope traits derived from milk, fat, and protein yield records are referred to as heat tolerance milk (**HTMYslope**), fat (**HTFYslope**), and protein (**HTPYslope**).

Genotypes

Two genotype datasets were prepared for the above cows and bulls with heat tolerance phenotypes: 50k SNP chip and 15,098,486 imputed whole-genome sequence variants (**WGS**).

The WGS was imputed using the genomic sequence data from the Run7 of the 1000 Bull Genome Project based on the ARS-UCD1.2 reference genome (http://1000bullgenomes.com/), and variants were filtered on the estimated imputation accuracy (R^2 > 0.4) and minor allele frequency (MAF > 0.005). Detailed imputation procedure is described by [22].

Study design: Discovery, Reference, and Validation datasets

The animals with genotypes and heat tolerance phenotypes comprised Holsteins (29,107 ♀/3,323 ♂), Jersey (6,338 ♀/1,364 ♂), and Holstein-Jersey crossbreds (790 ♀/0 ♂). These animals were split into 3 independent groups to achieve the specific objectives: 1) QTL discovery set – used to discover informative sequence markers for heat tolerance 2) Reference set – used to develop genomic prediction equations 3) independent validation sets – used to assess genomic prediction accuracy. The validation sets included three independent breed subsets: Holstein, Jersey, and crossbred cows. Across all the prediction scenarios, we ensured that the QTL discovery set used in GWAS was independent of the reference set used in genomic predictions to minimise bias in the predictions [21]. The different sets of animals used for each group (QTL discovery, reference, and validation) are described schematically in Figure 1, with a more detailed description as follows:

Scenario 1: this was aimed at testing the value of pre-selected sequence variants from Holsteins in the genomic prediction of the same breed as well as in the prediction of other numerically smaller breeds, including Jersey and crossbred cows. 1) QTL discovery set – comprised 20,623 Holstein cows born in 2012 or earlier 2) Reference set – comprised 3,323 Holstein bulls. None of these bulls sired cows in the discovery set to ensure the independence of the phenotypes between the two datasets 3) Validation sets – a) comprised of 1,223 younger Holstein cows (born in 2013 or later) that were not daughters of the Holstein bulls used in the reference set b) Jersey (N = 6,338 ♀) c) crossbreds (N = 790 ♀). Each of the three validation sets was randomly split into two subsets of approximately equal size (Supplementary Table S2) to facilitate the calculation of standard errors of prediction.

Scenario 2: we tested the hypothesis that using pre-selected informative markers from a multi-breed population improves the accuracy of predictions compared to pre-selected markers from the single-breed QTL discovery set. 1) QTL discovery set – we combined subsets of older cows that were: Holstein (N = 20,623 ♀; born in 2012 or earlier) and Jersey (N = 5,143 ♀; calved for the first time in 2014 or earlier); 2) Reference set – Holstein bulls (N = 3,323); 3) Validation sets – a) Holsteins (N = 1,223 ♀; as described above for 'scenario 1' above) b) Jersey (N =

1,195; younger cows that calved for the first time in 2014 or later) 3) crossbred cows (N = 790; as described for 'scenario 1' above). As in 'scenario 1' above, each validation set was randomly split into two subsets of approximately equal (Supplementary Table S2). The subsets for Holsteins and crossbred validation sets were the same as in 'scenario 1'.

Scenario 3: we tested the accuracy of prediction when using a multi-breed reference set as follows: 1) QTL discovery set – Holstein cows (N = 20,623; born in 2012 or earlier as described for 'scenario 1' above, i.e., single-breed discovery set); 2) Reference set – combined (multi-breed) set of bulls for Holsteins (N = 3,323 ♂; as described for 'scenario 1' above) and Jersey (N = 852 ♂); 3) Validation sets – a) Holstein cows (N = 1,223; as described for 'scenario 1 and 2' above) b) Jersey cows (N = 431) that were not daughters of the bulls used in the multi-breed reference set c) crossbred cows (N = 790; as described for 'scenarios 1 and 2' above). Validation sets were split as described for 'scenario 1 and 2' above. The subsets for Holsteins and crossbred validation sets were the same as in 'scenarios 1 and 2'.

QTL discovery and selection of informative markers ('top SNPs')

To identify informative sequence variants for heat tolerance traits (using the "Discovery" sets described above), we performed a genome-wide association study (GWAS) using mixed linear models to test associations between individual SNP and cow's slope traits using GCTA software [24]. The details of the GWAS for the Holstein discovery set are described by [22]. Briefly, a linear model was fitted to each cow's (N = 20,623 Holsteins) slopes for production trait [HTMYslope, HTFYslope, and HTPYslope] (pre-adjusted for the nongenetic effects described by [23]), for each autosomal SNP (~ 15 million SNPs). The model included a genomic relationship matrix (GRM) constructed from 50k SNP genotype data of the cows. The same model was used when performing GWAS for the multi-breed (Holstein and Jersey cows; N = 25,766) QTL discovery set except that an additional binary covariate was fitted to account for the breed effect.

To increase the power of GWAS to identify pleiotropic variants for heat tolerance from the three slope traits, we combined the above single-trait GWAS results in a multi-trait meta-GWAS (following methods described by [25]) and described for the Holstein data set in [22].

Using either the single-trait or multi-trait GWAS results, we selected informative variants defined as 'top SNPs' for each slope trait as follows:

1. We chose the most significant SNP from within each 100 kb window and then in each sliding 50 kb window along each chromosome. To be selected, the SNP had to pass

either a more stringent GWAS threshold of $-log_{10}(p\ value) \geq 3$ or a relaxed GWAS cut-off of $-log_{10}(p\ value) \geq 2$. This relaxed GWAS cut-off is suited for capturing variants with small to large effect sizes for heat tolerance [22].

2. Among each set of selected 'top SNPs', we removed one SNP of any pair in strong LD ($r^2 > 0.95$) using PLINK software [26], with the [−indep-pairwise 50 5 0.95] option, where LD is calculated within 50 SNPs sliding window, each time sliding five SNPs along the chromosome.

Genomic prediction

We used BayesR [16, 27] to estimate prediction accuracies using the 50k SNP panel and compared this to accuracies estimated by adding pre-selected 'top SNPs' to the 50k SNP set (i.e., 50k + 'top SNPs') obtained from the BayesRC method (MacLeod et al., 2016). The Australian dairy industry currently uses the 50k SNP panel for routine genomic evaluations; thus, it served as the standard to test the added value of selected sequence variants ('top SNPs'). Furthermore, the industry 50k SNP panel includes a set of variants that were not selected intentionally for heat tolerance; thus, this was ideal for our study.

The BayesR model fitted to the reference bulls (N = 3,323) for 42,572 variants (with MAF > 0.005) from 50k SNP panel was, $y = X\beta + Wv + e$, where $\mathbf{y}$ = vector of heat tolerance slope phenotypes (HTMYslope, HTFYslope, and HTPYslope); $\mathbf{X}$ = design matrix allocating phenotypes to the fixed effects, where fixed effects included the overall mean and a dummy random categorical variable with values 0 and 1 (this dummy variable was required as a placeholder for our inhouse BayesR program to run); $\boldsymbol{\beta}$ = vector of fixed effect solutions; $\mathbf{W}$ = centred design matrix of SNP genotypes; $\mathbf{v}$ = vector of SNP effects, modelled to have four possible normal distributions corresponding to zero-, small-, medium- and large-sized effect, respectively; $\mathbf{e}$ = vector of residual errors $N(0, \mathbf{E}\sigma_e^2)$, where $\mathbf{E}$ is a diagonal matrix calculated as $diag(1/w_i)$, with w_i being a weighting factor for ith animal calculated differently for cows and bulls based on the available number records following Garrick et al. (2009) assuming that 0.2 of the genetic variance is not accounted by the SNPs. The same model was used when analysing the multi-breed reference population (Holstein and Jersey; N = 4,175), except that a binary covariate was fitted to account for the breed effect. To account for polygenic effects, we tested models with or without pedigree relationships, which yielded correlation estimates of SNP effects of around 1.0. Therefore, based on these preliminary analyses, we decided not to include pedigree data in subsequent models.

We used the BayesRC method [28] to analyse the 50k + 'top SNPs' dataset. The BayesRC allows pre-allocation of variants to 2 or more classes assuming a different posterior mixture distribution within each class if the class is enriched for informative SNPs. In our case, the SNPs from the 50k array (42,572) were allocated to class I and the selected 'top SNPs' to a separate class II, because the latter may be enriched with causal and/or highly predictive mutations for heat tolerance. For both BayesR and BayesRC models, we performed five MCMC replicate chains, each with 40,000 iterations of which 20,000 were discarded as burn-in for all the traits. These iterations gave stable convergence across the 5 replicates. To facilitate the calculation of standard errors, and based on the number of animals available, we performed all analyses for two subsets of approximately equal size for each validation set (Supplementary Table S2).

For each analysis (described above), the accuracy of prediction was calculated as described in [11]: $Accuracy = \frac{r_{GBV,phen}}{\sqrt{h^2}}$, where $r_{GBV}, phen$ = correlation of GBV and TD phenotypes (slope traits); (h^2 = genomic heritability of the trait computed from 50k SNP data based on 29,107 Holstein cows), as described earlier (Supplementary Table S1). The corresponding standard errors of the accuracies were estimated as: $SE = SD/\sqrt{N}$, where N = number of random validation subsets (N = 2); SD = standard deviation of the accuracies of prediction calculated from the validation sets. The bias of prediction accuracies for different traits were assessed as the regression coefficient of the TD phenotypes on the GBV in the validation set and their corresponding standard errors calculated as described for the SE of the accuracies of prediction above.

Results

Genomic heritability

Genomic heritability estimates based on 29,107 Holstein cows using 50k SNP array were similar for all the slope (heat tolerance) traits (Supplementary Table S1). The genomic heritability estimates based on Jersey cows (N = 6,338) were comparable to those for Holstein cows with values of 0.26 ± 0.02, 0.23 ± 0.02, and 0.25 ± 0.02 for HTMYslope, HTFYslope and HTPYslope traits, respectively (Supplementary Table S1). However, the values for crossbred cows (N = 790) were estimated with large standard errors [0.58 ± 0.10 (HTMYslope); 0.34 ± 0.11 (HTFYslope); 0.51 ± 0.10 (HTPYslope)], most likely due to the small sample size used. In this study, we computed the accuracy of genomic predictions across

all validation sets using the heritability estimates from Holstein cows (N = 29,107) that were estimated with the smallest standard errors.

Genomic prediction using single-breed QTL discovery set ('scenario 1')

Table 1 includes the number of selected informative sequence variants for heat tolerance defined as 'top SNPs' from single-trait GWAS and multi-trait meta-analyses of Holstein cow discovery set (i.e., single-breed discovery set; see methods section – 'scenario 1'). Using a more stringent GWAS cut-off threshold of -log10(p-value) ≥ 3 resulted in about 5-fold lower number of selected 'top SNPs' than a comparatively relaxed GWAS cut-off of -log10(p-value) ≥ 2. The number of selected 'top SNPs' at -log10(p-value) ≥ 2 from single-trait GWAS (after pruning pairs of markers in strong LD, $r^2 > 0.95$) were 9,481 ± 244 and those selected at -log10(p-value) ≥ 3 were 1,758 ± 117 (Table 1). The largest number of 'top SNPs' were selected for HTPYslope, followed by HTFYslope and HTMYslope (Table 1). Although the number of variants that passed the GWAS cut-off was greatest for HTPYslope, the strength of the GWAS signal (peak) across the genome was relatively weak for this trait compared to the other traits (i.e., HTMYslope and HTFYslope). A large proportion (> 50%) of the selected 'top SNPs' have lower MAF compared to the SNPs in the 50k panel (Supplementary Figure S1). Compared to single-trait GWAS, and as expected, the number of selected 'top SNPs' were generally higher for multi-trait meta-analysis of slope traits at a more stringent [-log10(p-value) ≥ 3; N = 2,365 SNPs] and at a relaxed [-log10(p-value) ≥ 2; N = 9,090 SNPs] GWAS cut-off thresholds (Table 1).

Table 1 Number of informative markers for heat tolerance defined as 'top SNPs' selected from single-trait GWAS and multi-trait meta analyses of heat tolerance slope traits of Holstein discovery cow set (N = 20,623).

	Single-trait GWAS		Multi-trait meta-analysis	
Trait	Top SNPs (logP = 2)	Top SNPs (logP = 3)	Top SNPs (logP = 2)	Top SNPs (logP = 3)
HTMYslope	9,207 (51,750)	1,654 (44,219)	9,090 (51,636)	2,365 (44,929)
HTFYslope	9,352 (51,894)	1,708 (44,277)	9,090 (51,636)	2,365 (44,929)
HTPYslope	9,633 (52,168)	1,624 (44,190)	9,090 (51,636)	2,365 (44,929)

Markers were selected based on the GWAS cut-off thresholds of -log10(p-value) ≥ 2 and -log10(p-value) ≥ 3; The values in brackets are the final number of SNPs after adding selected 'top SNPs' to the 50k SNP data used in the BayesRC analyses (i.e., 42,572 SNPs + top SNPs); Traits are defined as heat tolerance milk (HTMYslope), fat (HTFYslope) and protein (HTPYslope) yield slope traits.

Figure 2 shows the accuracy of predictions when the selected 'top SNPs' from a single-breed (Holstein cows; N = 20,623) QTL discovery set were added to the standard 50k SNP array and analysed using the BayesRC models. For this comparison, the reference set was only Holstein bulls (N = 3,323) and the validation set was Holsteins (N = 1,223), Jersey (N = 6,338) and crossbred (N = 790) cows. The gain in accuracy for different traits and models varied across the three validation sets. The increase in the prediction accuracy was somewhat consistent for the HTMYslope trait across most of the different cases (50k + 'top SNPs') tested, but not for HTFYslope and HTPYslope trait, particularly in the Jersey validation set. In general, the increase in accuracy ranged from 0.001 to 0.09, with the largest estimate (0.09) observed for HTMYslope in the crossbred validation set.

In most cases, except in Jerseys, the bias of prediction (assessed as the regression coefficient of the slope phenotypes on the GBV in the validation sets) was > 1.0 (Supplementary Figure S4), indicating 'deflation' or under prediction, meaning less variance among predicted than the observed values. We observed the most extreme bias (>1.7) for HTMYslope in the crossbreds (N = 790) and Jersey (N = 431) validation sets (bias < 0.5), likely due to the small sample size and population used. The prediction bias was even more pronounced when the selected 'top SNPs' were added to the 50k SNP data in all BayesRC models compared to the estimates from the BayesR using only 50k SNP data. Notably, the bias of prediction was generally lower (values closer to 1.0) for the intercept traits (represent the level of milk production) when compared to heat tolerance traits (Supplementary Figure S3 and S4).

The 'top SNPs' selected from the relaxed GWAS cut-off value of -log10(p-value) ≥ 2 (~9,000 SNPs) yielded, in general, a greater lift in accuracy compared to prediction estimates based on the 'top SNPs' from a more stringent GWAS threshold (-log10(p-value) ≥ 3; ~2,000 SNPs) across most traits and validation sets (Figure 2). On average, using 'top SNPs' from the relaxed GWAS cut-off (-log10(p-value) ≥ 2) resulted in a 2% gain in accuracy when compared to selected 'top SNPs' based on a more stringent GWAS threshold (-log10(p-value) ≥ 3). In general, using 'top SNPs' from the more stringent GWAS cut-off yielded more bias of prediction than the 'top SNPs' from a relaxed GWAS cut-off threshold.

The 'top SNPs' from single-trait GWAS yielded (on average) about 1% greater gain in accuracy compared to the 'top SNPs' from multi-trait meta-GWAS of slope traits. The increase in accuracy based on 'top SNPs' from single-trait GWAS ranged from 0.001 (HTFYslope) to 0.09 (HTMYslope) both in the crossbreds. In contrast, the average lift in accuracy based on the

selected 'top SNPs' from the meta-analysis of slope traits was 0.03 with values ranging from 0.009 (HTPYslope) to 0.07 (HTMYslope), both in the crossbred cows. Overall, the above results show that the 'top SNPs' from single-trait GWAS at the relaxed cut-off threshold (-log10(p-value) ≥ 2) resulted in a more lift in accuracy than the 'top SNPs' from stringent or those from multi-trait meta-GWAS. Therefore, and hereafter, we only report the results based on the selected 'top SNPs' from the single-trait GWAS at the relaxed cut-off threshold.

Notably, the prediction accuracy decreased considerably for HTFYslope and HTPYslope traits when the selected 'top SNPs' from Holsteins cows were used in Jersey with a more decrease in accuracy when using 'top SNPs' from single-trait GWAS than those from the multi-trait meta-analysis (Figure 2). We observed the largest drop in accuracy for HTFYslope (10%) and HTPYslope (9%) traits for Jerseys when using selected 'top SNPs' from single-trait GWAS at the stringent GWAS cut-off (-log10(p-value) ≥ 3). We also observed a slight drop in accuracy when the selected 'top SNPs' from the Holstein cow discovery set were used in the crossbreds in some prediction scenarios (Figure 2).

To test whether allocating selected informative markers to a separate class (see methods) in the BayesRC can show added benefit in our study, we combined 50k + selected 'top SNPs' from single-breed (Holsteins) QTL discovery set and modelled using the BayesR – a method which does not allow defining priors [22]. This test was performed only for two traits in the Holstein validation set. The BayesRC gave 2% (HTMYslope) and 1% (HTFYslope) more lift in the accuracy, indicating its superiority compared to the BayesR method.

Genomic prediction using multi-breed QTL discovery set ('scenario 2')

When Holstein cows (N = 20,623) were combined with Jersey cows (N = 5,143) in the QTL discovery set (i.e., multi-breed QTL discovery set; see methods section – 'scenario 2'), we found a lower number of selected 'top SNPs' (after pruning pairs of markers in strong LD, r^2 > 0.95) from single-trait GWAS at -log10(p-value) ≥ 2 [HTMYslope = 6,132; HTFYslope = 6,286; HTPYslope = 6,422] compared to those from single-breed QTL discovery set at the same significance cut-off as described above.

Figure 3 shows the gain in accuracy of prediction when the selected 'top SNPs' (GWAS cut-off of -log10(p-value) ≥ 2) from multi-breed (Holstein + Jersey cows) QTL discovery set were added to the 50k SNP data in which the reference set was only Holstein bulls. The change in accuracy across all traits and validation sets ranged from -0.05 (HTPYslope) in Jersey to 0.11 (HTMYslope) in crossbred cows. In the Holstein validation set (N = 1,223), the accuracy of

prediction increased across all traits with the greatest gain for HTPYslope (0.03) followed by HTFYslope (0.02) and HTMYslope (0.01), respectively. In this validation set, the bias was > 1.0 across all the traits, indicating under prediction. The bias decreased slightly for HTMYslope but increased for HTPYslope and HTFYslope traits when the 'top SNPs' were fitted in the BayesRC method (Figure 3).

In the Jersey validation set (N = 1,195), the change in accuracy was inconsistent across traits (Figure 3). When using the selected 'top SNPs' from the multi-breed QTL discovery set, the accuracy increased for HTMYslope (0.03) and HTFYslope (0.02) but decreased for HTPYslope (-0.05). These values contrast with those obtained from using selected 'top SNPs' from the single-breed QTL discovery set (only Holsteins; see methods 'scenario 1'), where we found a change in accuracy of 0.09, 0.04, and 0.01 for HTMYslope, HTFYslope, and HTPYslope, respectively, when using a smaller subset of Jersey cows (i.e., N = 1,195) instead of 6,338 cows (as in the 'scenario 1'). In the 'scenario 2' analyses, the prediction estimates for the HTFYslope trait based on the 50k SNP panel were 'inflated' or over predicted (bias < 1.0), whereas the HTMYslope and HTPYslope were over predicted (Figure 3). In the BayesRC, the bias for HTPYslope changed from 1.11 (50k) to 0.78 (BayesRC), indicating more biased (under-prediction) prediction estimates when fitting the 'top SNPs' for this trait.

In the crossbreds (N = 790), using 'top SNPs' discovered in the multi-breed (Holsteins + Jersey cows) set led to a change in accuracy of 0.11, -0.005, and -0.03 for HTMYslope, HTFYslope, and HTPYslope, respectively. In contrast, using 'top SNPs' from single-breed (only Holsteins) QTL discovery set yielded a change in accuracy of 0.09, 0.02, and -0.006 for HTMYslope, HTFYslope, and HTPYslope, respectively. The bias for HTMYslope was extreme (> 1.7) compared to the other traits. In the crossbred validation set ('scenario 2'), the bias increased more for HTMYslope but decreased for HTFYslope and HTPYslope when fitting the selected 'top SNPs' in the BayesRC (Figure 3).

Genomic prediction using multi-breed reference set ('scenario 3')

When we used a multi-breed (Holstein + Jersey bulls) reference set in which the 'top SNPs' were from only Holstein cow QTL discovery set (single-breed; see methods section – 'scenario 3'), we found a consistent lift in the accuracy of prediction in most cases (Figure 4). The accuracy of prediction decreased only for HTFYslope (-0.06) and HTPYslope (-0.002) in the Jersey validation set for this scenario. The change in accuracy ranged from [0.04 to 0.05], [-0.06 to 0.04], and [0.04 to 0.10] in the Holsteins (N = 1,223), Jersey (N = 431) and crossbred

(N = 790) cow validation set, respectively (Figure 4). These changes in accuracy are higher compared to those found when using single-breed reference set (Figure 2; 'scenario 1') with values ranging from [-0.01 to 0.06], [-0.10 to 0.05], [-0.06 to 0.09] for Holsteins (N = 1,223), Jersey (N = 6,338), and crossbred (N = 790) validation set. To be more comparable, when considering only a subset of Jersey cows (N = 431) in the validation set where the reference set is single-breed (only Holstein bulls; 'scenario 1'), we found a change in accuracy of -0.02, 0.03, and -0.06 for HTMYslope, HTFYslope, and HTPYslope traits, respectively. Compared to estimates from the 'scenario 1' and 'scenario 2' analyses above, we observed the lowest bias (i.e., values around 1.0) when using the multi-breed reference set in the Holstein validation set. However, in the Jersey validation set, we found extreme bias (> 2.0) for HTPYslope, whereas the bias was small for HTFYslope. In the crossbreds, the bias was high for HTMYslope (> 1.5) and HTFYslope (> 1.3), whereas we observed a small bias (values closer to 1.0) for the HTPYslope trait.

Discussion

In this study, we present a genomic prediction analysis of heat tolerance traits using a large sample size of over 40,000 cattle, comprising Holsteins, Jersey, and crossbreds. The primary objective was to investigate if selected sequence variants from a GWAS of Holsteins benefits genomic prediction of heat tolerance phenotypes in the same breed (i.e., within-breed prediction). The hypothesis is that the selected variants are linked to causal mutations underpinning the genetic basis for heat tolerance; therefore, could enable more accurate and sustained genomic selection for this trait. In addition, we also tested the value of pre-selected variants from Holsteins for the genomic prediction of breeds with numerically smaller sample sizes, such as Jersey and crossbreds. Furthermore, we investigated the benefits of using informative markers from multi-breed (Holstein + Jersey cows) QTL discovery set in the genomic predictions of heat tolerance. Overall, our results show that we can increase the prediction accuracy of heat tolerance by up to 10% in some scenarios when pre-selected sequence variants are added to the 50k SNP panel.

We used BayesR and BayesRC methods to test different prediction scenarios. For BayesR, using only 50k SNP data, we found high accuracies of prediction in Holsteins and crossbreds compared to Jersey cows. We expected a lower accuracy in Jerseys because we used Holstein bulls as a reference set for genomic predictions. These breeds are genetically divergent and may have different linkage disequilibrium of variants with causal mutations, may not share all the same causal variants, or some variant effects may differ between these breeds [29]. As such,

when we combined Holstein and Jersey bulls in the reference set (multi-breed reference set) and performed analysis using BayesR (without pre-selected 'top SNPs'), we found a substantial improvement in the accuracy of prediction across all traits for Jerseys which is consistent with the multi-breed genomic predictions reported in previous studies [e.g., 16, 29].

For the BayesRC models, where 50k + selected 'top SNPs' was fitted in the analysis, we demonstrated a consistent increase in the accuracy across traits when the 'top SNPs' that were selected from Holsteins and used in the genomic prediction of the same breed (i.e., within breed QTL discovery and validation set). Similarly, using 'top SNPs' from Holstein discovery set in crossbred cattle based on the BayesRC performed reasonably well, which we expected since our crossbred cows have substantial Holstein genes (i.e., there were mostly HHHJ or HHJJ crosses). The gain in accuracy of prediction for Holsteins and crossbreds likely benefited, in part, from a powerful GWAS QTL discovery (we used a sample size of 20,623 Holstein cows, each having around 15 million imputed sequence variants) and the methodology used for genomic prediction. To date, comparable GWAS have used a sample size of at most around 5,000 [e.g., 5] to search for variants associated with heat tolerance in dairy. We expect an even more increase in accuracy of prediction in the future with larger sample sizes for GWAS to increase the power of QTL discovery.

On the other hand, the genomic predictions performed somewhat poorly in Jerseys, particularly for HTFYslope and HTPYslope traits, where the accuracies decreased when the selected 'top SNPs' from the Holstein discovery set were added to 50k SNP set and used in the BayesRC. Given that Holsteins and Jersey are genetically divergent breeds, using informative QTLs from Holstein in Jersey validation may have introduced noise into genomic predictions since the common QTLs may not be tracked across these breeds, leading to the observed drop in the accuracy. However, it is unclear why the accuracy increased for HTMYslope in Jerseys but not for HTFYslope and HTPYslope. One possible reason could be due to the different genetic architecture of these traits. This is evidenced by the smaller number of 'top SNPs' for HTMYslope detected from GWAS of Holsteins at the relaxed cut off ($p < 0.01$) (Table 1) compared to HTPYslope and HTFYslope traits, suggesting that HTMYslope is controlled by relatively few QTLs with large effects compared to HTPYslope and HTFYslope. By comparing the GWAS of Holsteins (N = 20,623) and Jerseys (N = 6,338) cows, we found the greatest overlap of top significant SNPs that were at least within 1 Mb regions for HTMYslope mapping to the genomic regions showing strong GWAS signal in chromosome 5, 14, 20, and 25. This overlap, in part, explains the greater consistency of increase in accuracy for HTMYslope

494 compared to HTFYslope and HTPYslope traits. In this context, our findings are in line with
495 [27], who reported that only a fraction of QTLs for milk yield segregate across Holsteins and
496 Jersey cattle. Overall, the results suggest that the informative markers from the Holsteins are
497 of little or no value to the prediction of Jersey breeds. In addition, it appears that HTMYslope
498 could be a more reliable indicator trait of heat tolerance and could be given greater weight in
499 the selection index. Australian dairy industry currently gives more economic weight to
500 HTPYslope (6.92) than HTMYslope (-0.10) or HTFYslope (1.79) slope traits in the calculation
501 of heat tolerance genomic breeding values based on weights for milk production traits [30, 31].

502 Previous research studies in cattle [e.g., 17, 18] have reported a more precise mapping of
503 putative causative mutations when using multi-breed populations in GWAS and have
504 implicated pathways underpinning heat tolerance [22]. In this study, we found some
505 improvement in the predictions, especially in Jersey, when using 'top SNPs' from a discovery
506 set of combined Holstein and Jersey cows (i.e., multi-breed QTL discovery set). For example,
507 the accuracy increased by 3% for HTFYslope when using 'top SNPs' selected from multi-
508 breed discovery set in Jersey compared to a drop of 6% when the 'top SNPs' from Holstein
509 QTL discovery set (single-breed) was used in the BayesRC (Figure 3). In principle, combining
510 divergent breeds in the QTL discovery set may help to break long-range LD, such that the
511 selected 'top SNPs' are closer to the causal mutations [16] compared to using a single-breed
512 QTL discovery set. For example, the top significant SNP on chromosome 14 mapped to the
513 upstream region of SLC52A2 and intronic to HSF1 gene in single-breed and multi-breed QTL
514 discovery set, respectively (Supplementary Figure S5). The latter gene has been linked to
515 thermotolerance in dairy cattle [5, 6, 22]. The smaller number of 'top SNPs' detected from
516 multi-breed than within-breed QTL discovery set in our study is consistent with the work of
517 [18] attributed, in part, because not all causal variants segregate across Holstein and Jersey
518 breeds.

519 However, we could still see a decrease in accuracy (-5%) for HTPYslope when using 'top
520 SNPs' from a multi-breed discovery set in Jersey, although not as high as (-8%) found when
521 using 'top SNPs' from single-breed (Holsteins) discovery set. Notably, our multi-breed QTL
522 discovery set was highly dominated by Holsteins which, in part, explains the limited gain in
523 accuracy when the selected 'top SNPs' from the multi-breed discovery set were used in the
524 Jerseys. Besides, we used Holstein bulls as a reference set in genomic predictions in Jerseys.
525 Since these breeds are divergent, a better approach to improve genomic predictions in Jerseys
526 would have been to use 'top SNPs' from multi-breed or within-breed (Jersey) and a reference

set of the same breed (Jersey) or multi-breed set. However, compared to Holsteins, the numerically smaller number of Jersey animals meant that it was not possible to split Jersey dataset in our study to obtain independent subsets with sufficient power for use in the QTL discovery and reference set for genomic predictions. This implies that there may be more room for improvement in accuracy for Jerseys when more animals are genotyped in the future.

We compared the added value of informative markers (i.e., 'top SNPs') from single-trait GWAS versus multi-trait meta-GWAS in the genomic predictions. The meta-analysis of slopes aimed to increase the power of GWAS and obtain a set of 'top SNPs' with putative pleiotropic effects for heat tolerance phenotypes. There is a recent trend towards developing custom SNP arrays that include variants with pleiotropic effects across multiple traits [32, 33]. In this study, we found comparable gain in accuracy when using 'top SNPs' from single-trait GWAS or meta-analysis. However, the gain in accuracy was, on average, ~1% lower when using 'top SNPs' from meta-GWAS than those from single-trait GWAS, although the accuracies varied considerably across traits and validation set used (Figure 3). Our recent work [22] suggests that heat tolerance traits (milk, fat, and protein slopes) are regulated somewhat differently in heat-stressed dairy cows. As such, we think that the relatively lower accuracy realized from using selected 'top SNPs' from the meta-GWAS of slope traits could be due to the possible inclusion of non-causal 'top SNPs' in genomic prediction, which arose from combining SNP effects for different heat tolerance phenotypes.

In general, we have demonstrated a lift in the accuracy of heat tolerance when informative sequence markers are added to 50k SNP panel by up to 7%, 5%, and 10% in Holsteins, Jersey, and crossbred cows, respectively. Our findings are within the range of those reported for complex traits in cattle [e.g., 34] and sheep [e.g., 11, 13]. For example, [13] reported an increase in accuracy by 9% for parasitic resistance in Australian sheep, while [34] found an increase of up to 6% for carcass traits in cattle. These results indicate that informative markers can be prioritised, especially in the development of customized SNP arrays [33]. Adding informative variants for heat tolerance to the custom SNP panels as in [33] ensures that higher accuracies are achieved, which will help to drive genetic gain for this trait. Moreover, we expect that the genetic prediction of this trait would be sustained over generations when informative variants that are closer to the causal mutations are included in the custom SNP panels, as demonstrated by [35]. These authors found that using the custom XT_50k SNP panel, which contains prioritised sequence markers, gave a consistent and superior accuracy of predictions (compared

559 to standard SNP panels) in crossbred cows (crossbreds represents “more distant relationships
560 or many generations”).

561 Most of our accuracy estimates were under-predicted or ‘deflated’ (i.e., bias > 1.0, meaning
562 less variance among predicted than observed values). In all our analyses, we have used only
563 bulls in the reference population and only cows in the validation of genomic predictions. As
564 such, the lower variance of bull phenotypes resulting from averaging daughter slope solutions
565 (see methods), in part, explains the observed bias, especially in the Holstein cow validation set.
566 To test this, we split Holstein cows into reference (older cows) and independent validation
567 (young cows) sets. Consequently, we found a bias < 1.0, which supports our hypothesis. The
568 fact that the bias of prediction, in most cases, were more pronounced when the selected ‘top
569 SNPs’ were added to the 50k SNP array and analysed in the BayesRC is consistent with some
570 previous studies [e.g., 19, 20], likely due to a phenomenon often called the “Beavis effect”
571 [36], which comes from the overestimation of the effect size of the pre-selected variants. The
572 lower bias found when fitting the selected ‘top SNPs’ from the stringent GWAS cut-off than
573 the relaxed GWAS cut-off is somewhat inconsistent with [20], who reported more bias when
574 markers were strongly pre-selected. Here, we used the Bayesian approach (BayesRC), while
575 [20] applied GBLUP in their work. Notably, the magnitude of bias observed in this study may
576 not be a big issue in the routine genetic evaluations of heat tolerance, where genomic breeding
577 values are often calculated jointly for bull and cow phenotypes based on different weightings
578 according to the amount of information [7, 30].

579 In this study, we have investigated the utility of pre-selected sequence variants in the genomic
580 prediction of heat tolerance for milk production traits (milk, fat, and protein yield). It is also
581 worthwhile to investigate the added value of prioritised sequence variants for heat tolerance on
582 other traits that are affected by heat stress (e.g., fertility) because there are likely to be benefits
583 from achieving higher systemic heat tolerance across multiple traits. This added value could
584 be significant considering economic selection indices, e.g. for the Australian dairy industry,
585 are formulated to capture different aspects of farm profitability, including production, fertility,
586 health, functional, and type as well as feed efficiency traits [31]. Selecting for thermotolerance
587 would be advantageous if the desire is to simultaneously achieve an optimal level of heat
588 tolerance for multiple traits [23]. Therefore, further studies are needed to investigate the
589 benefits of sequence variants in improving heat tolerance with respect to other traits likely to
590 be affected by heat and humidity, such as fertility and health traits.

Conclusions

The results show that the accuracy of genomic prediction for heat-tolerance milk yield traits (milk, fat, and protein) can be improved by up to 10% when the selected sequence variants linked to heat tolerance are added to the 50k SNP panel. However, when predicting across breeds using informative sequence markers from the Holstein cow discovery set in the prediction for Jersey animals, the pre-selected variants did not improve the accuracy, especially for heat tolerance fat and protein yield traits. We observed improved predictions, particularly in the Jersey validation set when using pre-selected markers from the multi-breed (Holstein + Jersey cows) discovery and the multi-breed reference population. Prioritised sequence markers from single-trait GWAS yielded greater accuracy than those from the multi-trait meta-analysis of slope traits. Overall, the results show that sequence variants can be prioritised to improve the accuracy of heat tolerance and has direct application in the development of customized SNP arrays, and functionally implicate the genomic regions of the variants in heat tolerance mechanisms.

Ethics approval and consent to participate

The data used in this study is used for routine genetic evaluations by DataGene Ltd (Melbourne, Australia) and conforms with the Australian dairy industry guidelines for data collection from commercial dairy farms.

Availability of data and materials

DataGene (DataGene Ltd., Melbourne, Australia; https://datagene.com.au/) are the custodians of the raw phenotype and genotype data of Australian farm animals. Research-related requests for access to the data may be accommodated on a case-by-case basis.

Competing interests

The authors declare no competing interests.

Funding

The authors are grateful for the support by DairyBio (Melbourne, Australia), funded by Dairy Australia (Melbourne, Australia), the Gardiner Foundation (Melbourne, Australia), and

Agriculture Victoria Research (Melbourne, Australia). The funders had no role in study design, data collection and analysis, preparation of the manuscript or the decision to publish.

Authors' contributions

JEP, HM and IMM, conceived the study, designed, and supervised the analyses. IMM assisted in the preparation and imputation of genotype data. EKC performed association and genomic prediction analyses and wrote the first draft. All authors contributed to the formal data analysis, results interpretation, and discussions; and approved the final manuscript for publication.

Acknowledgements

We are grateful to the 1000 Bull Genomes Project consortium for providing access to the Run7 cattle sequence data (http://www.1000bullgenomes.com/). We thank Dr Bolormaa Sunduimijid (Agriculture Victoria Research) for the imputation of sequence data. We also thank Dr Ruidong Xiang (University of Melbourne) for the help in the data analysis. Thanks to DataGene Ltd. (Melbourne, Australia) and especially the participating farmers for providing the phenotype data.

References

[1] Polsky L., von Keyserlingk M.A., Invited review: Effects of heat stress on dairy cattle welfare, Journal of dairy science. 100 (2017) 8645-8657.
[2] St-Pierre N., Cobanov B., Schnitkey G., Economic losses from heat stress by US livestock industries, Journal of dairy science. 86 (2003) E52-E77.
[3] Carabaño M.J., Ramón M., Díaz C., Molina A., Pérez-Guzmán M.D., Serradilla J.M., Breeding and genetics symposium: breeding for resilience to heat stress effects in dairy ruminants. A comprehensive review, Journal of animal science. 95 (2017) 1813-1826.
[4] Ravagnolo O., Misztal I., Hoogenboom G., Genetic component of heat stress in dairy cattle, development of heat index function, Journal of dairy science. 83 (2000) 2120-2125.
[5] Wang T., Chen Y.-P.P., MacLeod I.M., Pryce J.E., Goddard M.E., Hayes B.J., Application of a Bayesian non-linear model hybrid scheme to sequence data for genomic prediction and QTL mapping, BMC genomics. 18 (2017) 618.
[6] Sigdel A., Abdollahi-Arpanahi R., Aguilar I., Peñagaricano F., Whole genome mapping reveals novel genes and pathways involved in milk production under heat stress in US Holstein cows, Frontiers in genetics. 10 (2019) 928.
[7] Nguyen T.T., Bowman P.J., Haile-Mariam M., Pryce J.E., Hayes B.J., Genomic selection for tolerance to heat stress in Australian dairy cattle, Journal of dairy science. 99 (2016) 2849-2862.

[8] Schaeffer L., Strategy for applying genome-wide selection in dairy cattle, Journal of animal Breeding and genetics. 123 (2006) 218-223.
[9] Schefers J.M., Weigel K.A., Genomic selection in dairy cattle: Integration of DNA testing into breeding programs, Animal Frontiers. 2 (2012) 4-9.
[10] Calus M.P., Bouwman A.C., Schrooten C., Veerkamp R.F., Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection, Genetics Selection Evolution. 48 (2016) 1-19.
[11] Moghaddar N., Khansefid M., van der Werf J.H., Bolormaa S., Duijvesteijn N., Clark S.A., Swan A.A., Daetwyler H.D., MacLeod I.M., Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations, Genetics Selection Evolution. 51 (2019) 72.
[12] Heidaritabar M., Calus M.P., Megens H.J., Vereijken A., Groenen M.A., Bastiaansen J.W., Accuracy of genomic prediction using imputed whole-genome sequence data in white layers, Journal of Animal Breeding and Genetics. 133 (2016) 167-179.
[13] Al Kalaldeh M., Gibson J., Duijvesteijn N., Daetwyler H.D., MacLeod I., Moghaddar N., Lee S.H., van der Werf J.H., Using imputed whole-genome sequence data to improve the accuracy of genomic prediction for parasite resistance in Australian sheep, Genetics Selection Evolution. 51 (2019) 32.
[14] van den Berg I., Boichard D., Guldbrandtsen B., Lund M.S., Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: a simulation study, G3: Genes, Genomes, Genetics. 6 (2016) 2553-2561.
[15] Brøndum R., Su G., Janss L., Sahana G., Guldbrandtsen B., Boichard D., Lund M., Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction, Journal of dairy science. 98 (2015) 4107-4116.
[16] Kemper K.E., Reich C.M., Bowman P.J., Vander Jagt C.J., Chamberlain A.J., Mason B.A., Hayes B.J., Goddard M.E., Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions, Genetics Selection Evolution. 47 (2015) 1-17.
[17] Sanchez M.-P., Govignon-Gion A., Croiseau P., Fritz S., Hozé C., Miranda G., Martin P., Barbat-Leterrier A., Letaïef R., Rocha D., Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle, Genetics Selection Evolution. 49 (2017) 1-16.
[18] Raven L.-A., Cocks B.G., Hayes B.J., Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle, BMC genomics. 15 (2014) 1-14.
[19] Raymond B., Bouwman A.C., Schrooten C., Houwing-Duistermaat J., Veerkamp R.F., Utility of whole-genome sequence data for across-breed genomic prediction, Genetics Selection Evolution. 50 (2018) 1-12.
[20] Veerkamp R.F., Bouwman A.C., Schrooten C., Calus M.P., Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle, Genetics Selection Evolution. 48 (2016) 95.
[21] MacLeod I., Bolormaa S., Schrooten C., Goddard M., Daetwyler H., Pitfalls of pre-selecting subsets of sequence variants for genomic prediction, in Proc Assoc Advmt Anim Breed Genet., 2017, Vol. 22, pp. 141-144.
[22] Cheruiyot E.K., Haile-Mariam M., Cocks B.G., MacLeod I.M., Xiang R., Pryce J.E., New loci and neuronal pathways for resilience to heat stress in animals, bioRxiv. (2021).
[23] Cheruiyot E.K., Nguyen T.T.T., Haile-Mariam M., Cocks B.G., Abdelsayed M., Pryce J.E., Genotype-by-environment (temperature-humidity) interaction of milk production traits in Australian Holstein cattle, J Dairy Sci. 103 (2020) 2460-2476.

[24] Yang J., Lee S.H., Goddard M.E., Visscher P.M., GCTA: a tool for genome-wide complex trait analysis, The American Journal of Human Genetics. 88 (2011) 76-82.
[25] Bolormaa S., Pryce J.E., Reverter A., Zhang Y., Barendse W., Kemper K., Tier B., Savin K., Hayes B.J., Goddard M.E., A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle, PLoS Genet. 10 (2014) e1004198.
[26] Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., De Bakker P.I., Daly M.J., PLINK: a tool set for whole-genome association and population-based linkage analyses, The American Journal of Human Genetics. 81 (2007) 559-575.
[27] Erbe M., Hayes B., Matukumalli L., Goswami S., Bowman P., Reich C., Mason B., Goddard M., Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels, Journal of dairy science. 95 (2012) 4114-4129.
[28] MacLeod I., Bowman P., Vander Jagt C., Haile-Mariam M., Kemper K., Chamberlain A., Schrooten C., Hayes B., Goddard M., Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits, BMC genomics. 17 (2016) 144.
[29] Hayes B.J., Bowman P.J., Chamberlain A.C., Verbyla K., Goddard M.E., Accuracy of genomic breeding values in multi-breed dairy cattle populations, Genetics Selection Evolution. 41 (2009) 1-9.
[30] Nguyen T.T., Bowman P.J., Haile-Mariam M., Nieuwhof G.J., Hayes B.J., Pryce J.E., Implementation of a breeding value for heat tolerance in Australian dairy cattle, Journal of dairy science. 100 (2017) 7362-7367.
[31] Byrne T., Santos B., Amer P., Martin-Collado D., Pryce J., Axford M., New breeding objectives and selection indices for the Australian dairy industry, Journal of dairy science. 99 (2016) 8146-8167.
[32] Xiang R., van den Berg I., MacLeod I.M., Daetwyler H.D., Goddard M.E., Effect direction meta-analysis of GWAS identifies extreme, prevalent and shared pleiotropy in a large mammal, Communications Biology. 3 (2020) 88.
[33] Xiang R., MacLeod I.M., Daetwyler H.D., de Jong G., O'Connor E., Schrooten C., Chamberlain A.J., Goddard M.E., Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations, Nature communications. 12 (2021) 1-13.
[34] de Las Heras-Saldana S., Lopez B.I., Moghaddar N., Park W., Park J.-e., Chung K.Y., Lim D., Lee S.H., Shin D., van der Werf J.H., Use of gene expression and whole-genome sequence information to improve the accuracy of genomic prediction for carcass traits in Hanwoo cattle, Genetics Selection Evolution. 52 (2020) 1-16.
[35] Khansefid M., Goddard M.E., Haile-Mariam M., Konstantinov K.V., Schrooten C., de Jong G., Jewell E.G., O'Connor E., Pryce J.E., Daetwyler H.D., Improving Genomic Prediction of Crossbred and Purebred Dairy Cattle, Frontiers in genetics. 11 (2020).
[36] Xu S., Theoretical basis of the Beavis effect, Genetics. 165 (2003) 2259-2268.

747 **Figures**

748 **Figure 1 Overview of the analyses describing three study scenarios.**

749 **Legend:** 'Scenario 1': QTL discovery set – comprised a subset of 20,623 older Holstein cows
750 (born 2012 or earlier); Reference set – comprised only Holstein bulls (N = 3,323) that were not
751 sires of cows in the discovery set; Validation sets – comprised Holsteins, Jersey and crossbred
752 cows. 'Scenario 2': QTL discovery set – comprised a combined set of Holstein (N = 20,623) +
753 Jersey cows (N = 5,143); Reference set - comprised only Holstein bulls (N = 3,323; as
754 described for 'scenario 1') that were not sires of the Holstein cows in the discovery set;
755 Validation sets – comprised Holstein (N = 1,223), Jersey (N = 6,338) and crossbred (N = 790)
756 cows. 'Scenario 3': QTL discovery set – comprised only Holstein cows (N = 20,623; as
757 described for 'scenario 1'); Reference set – comprised a combined set of Holstein (N = 3,323)
758 + Jersey (N = 852) bulls); Validation sets – comprised Holstein (N = 1,223), Jersey (N = 431)
759 and crossbred (N = 790) cows.

760 **Figure 2 Accuracy of genomic predictions (Holstein only reference) using either 50k SNP**
761 **data (colored grey) or 50k + a range of 'top SNPs' sets (selected from Holstein QTL**
762 **discovery set).**

763 **Legend:** The 'top SNPs' were selected from single-trait GWAS (colored blue) and multi-trait
764 meta-analysis (colored orange) at less stringent cut-off threshold of -log10(p-value) ≥ 2 [~9,000
765 SNPs] and at more stringent p-value of -log10(p-value) ≥ 3 [~2,000 SNPs]. Accuracy of
766 predictions are provided for 3 cow validation sets: Holsteins (**A;** N=1,223), Jersey (**B;**
767 N=6,338), and Holstein-Jersey crossbreds (**C;** N=790). Traits analysed are heat tolerance milk
768 (HTMYslope), fat (HTFYslope), and protein (HTPYslope) yield slopes. The genomic
769 predictions were generated using either BayesR (50K SNP set) or BayesRC (50K + top SNPs).
770 Vertical lines represent standard errors calculated from two random validation subsets.

Figure 3 Accuracy and bias of predictions in Holsteins (N = 1,223), Jersey (N = 1,195) and crossbred (N = 790) cows when using 50k + 'top SNPs' selected from multi-breed (Holstein + Jersey) QTL discovery set.

Legend: Holstein bulls (N = 3,323) were used as the reference set for genomic predictions. The 'top SNPs' were selected based on single-trait GWAS cut-off of [-log10(p-value) ≥ 2]. Traits analysed are heat tolerance milk (HTMYslope), fat (HTFYslope), and protein (HTPYslope) yield slopes. Vertical lines represent standard errors calculated from two random validation subsets.

Figure 4 Accuracy and bias of genomic predictions in Holsteins (N = 1,223), Jersey (N = 431) and crossbred (N = 790) cows when using multi-breed reference set (Holstein and Jersey bulls; N = 4,175).

Legend: The selected 'top SNPs' used in the BayesRC were from Holstein cow discovery set (N = 20,623) based on single-trait GWAS cut-off of [-log10(p-value) ≥ 2]. Traits analysed are heat tolerance milk (HTMYslope), fat (HTFYslope), and protein (HTPYslope) yield slopes. Vertical lines represent standard errors calculated from two random validation subsets.

Additional files

File name: Additional file 1

Format: .docx

Title of data: supplementary tables and figures

Figures

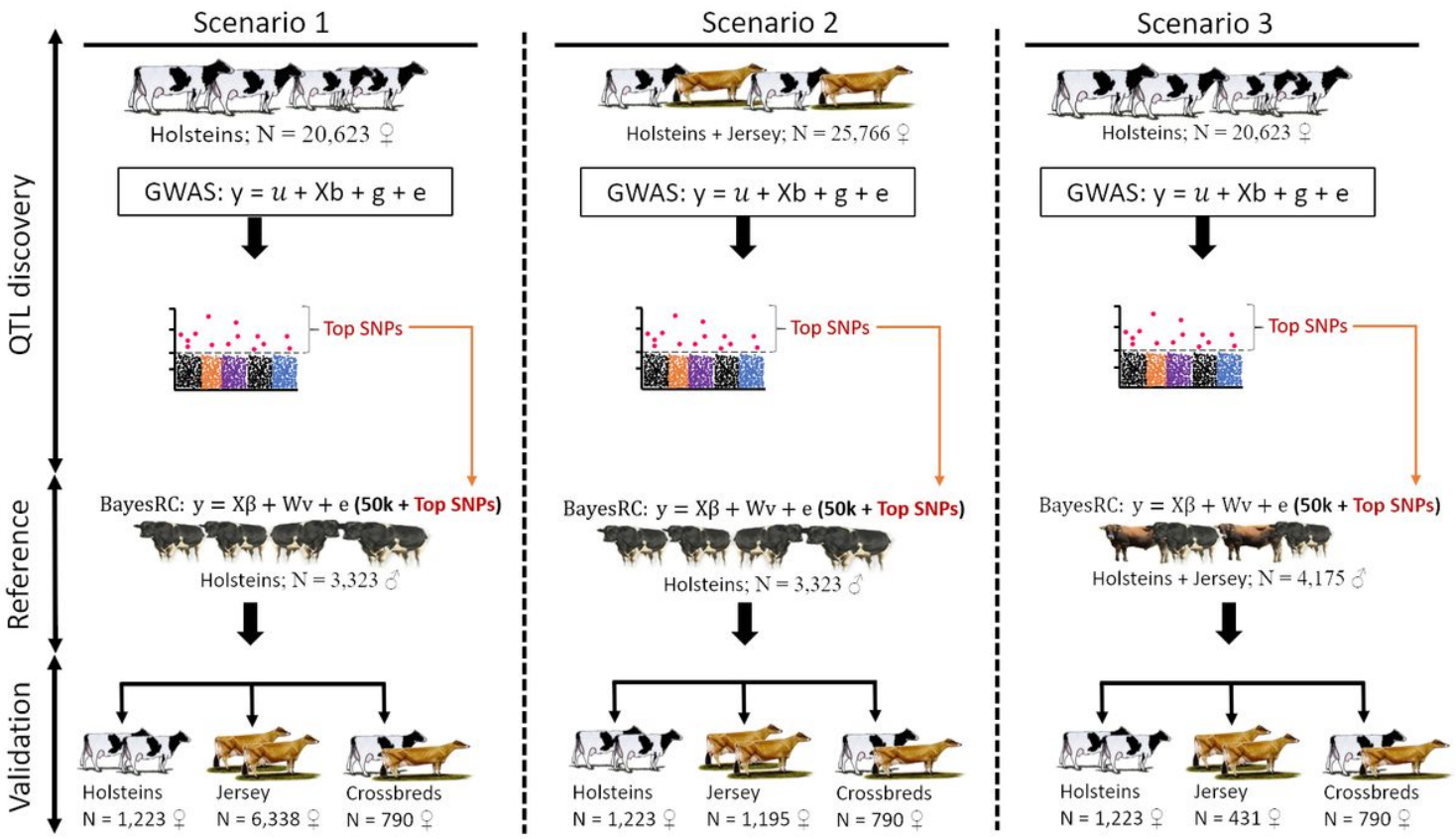

Figure 1

Overview of the analyses describing three study scenarios. 'Scenario 1': QTL discovery set – comprised a subset of 20,623 older Holstein cows (born 2012 or earlier); Reference set – comprised only Holstein bulls (N = 3,323) that were not sires of cows in the discovery set; Validation sets – comprised Holsteins, Jersey and crossbred cows. 'Scenario 2': QTL discovery set – comprised a combined set of Holstein (N = 20,623) + Jersey cows (N = 5,143); Reference set - comprised only Holstein bulls (N = 3,323; as described for 'scenario 1') that were not sires of the Holstein cows in the discovery set; Validation sets – comprised Holstein (N = 1,223), Jersey (N = 6,338) and crossbred (N = 790) cows. 'Scenario 3': QTL discovery set – comprised only Holstein cows (N = 20,623; as described for 'scenario 1'); Reference set – comprised a combined set of Holstein (N = 3,323) + Jersey (N = 852) bulls); Validation sets – comprised Holstein (N = 1,223), Jersey (N = 431) and crossbred (N = 790) cows.

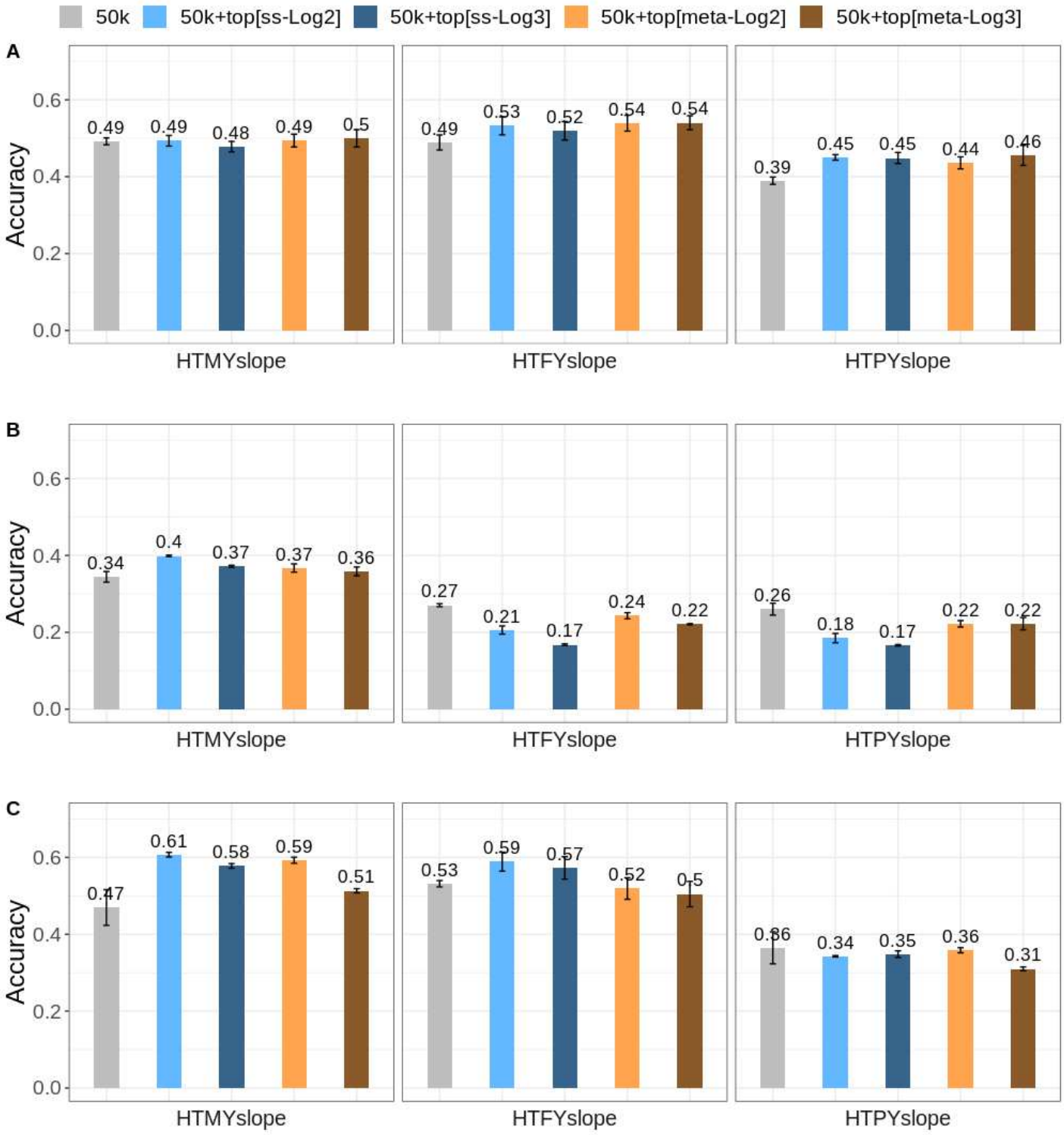

Figure 2

Accuracy of genomic predictions (Holstein only reference) using either 50k SNP data (colored grey) or 50k + a range of 'top SNPs' sets (selected from Holstein QTL discovery set). The 'top SNPs' were selected from single-trait GWAS (colored blue) and multi-trait meta-analysis (colored orange) at less stringent cut-off threshold of -log10(p-value) ≥ 2 [~9,000 SNPs] and at more stringent p-value of -log10(p-value) ≥ 3 [~2,000 SNPs]. Accuracy of predictions are provided for 3 cow validation sets: Holsteins (A; N=1,223),

Jersey (B; N=6,338), and Holstein-Jersey crossbreds (C; N=790). Traits analysed are heat tolerance milk (HTMYslope), fat (HTFYslope), and protein (HTPYslope) yield slopes. The genomic predictions were generated using either BayesR (50K SNP set) or BayesRC (50K + top SNPs). Vertical lines represent standard errors calculated from two random validation subsets.

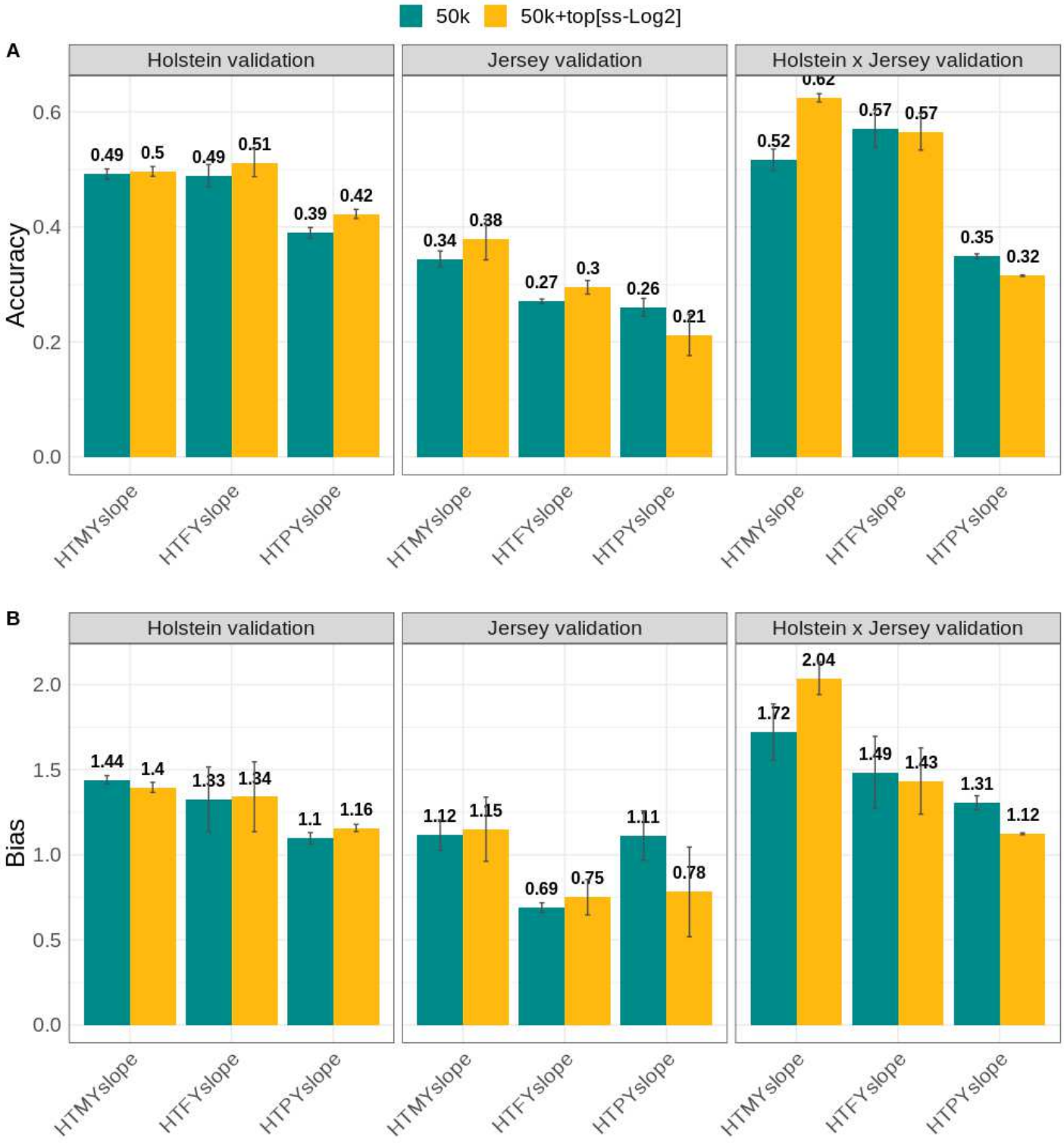

Figure 3

Accuracy and bias of predictions in Holsteins (N = 1,223), Jersey (N = 1,195) and crossbred (N = 790) cows when using 50k + 'top SNPs' selected from multi-breed (Holstein + Jersey) QTL discovery set. Holstein bulls (N = 3,323) were used as the reference set for genomic predictions. The 'top SNPs' were selected based on single-trait GWAS cut-off of [-log10(p-value) ≥ 2]. Traits analysed are heat tolerance milk (HTMYslope), fat (HTFYslope), and protein (HTPYslope) yield slopes. Vertical lines represent standard errors calculated from two random validation subsets.

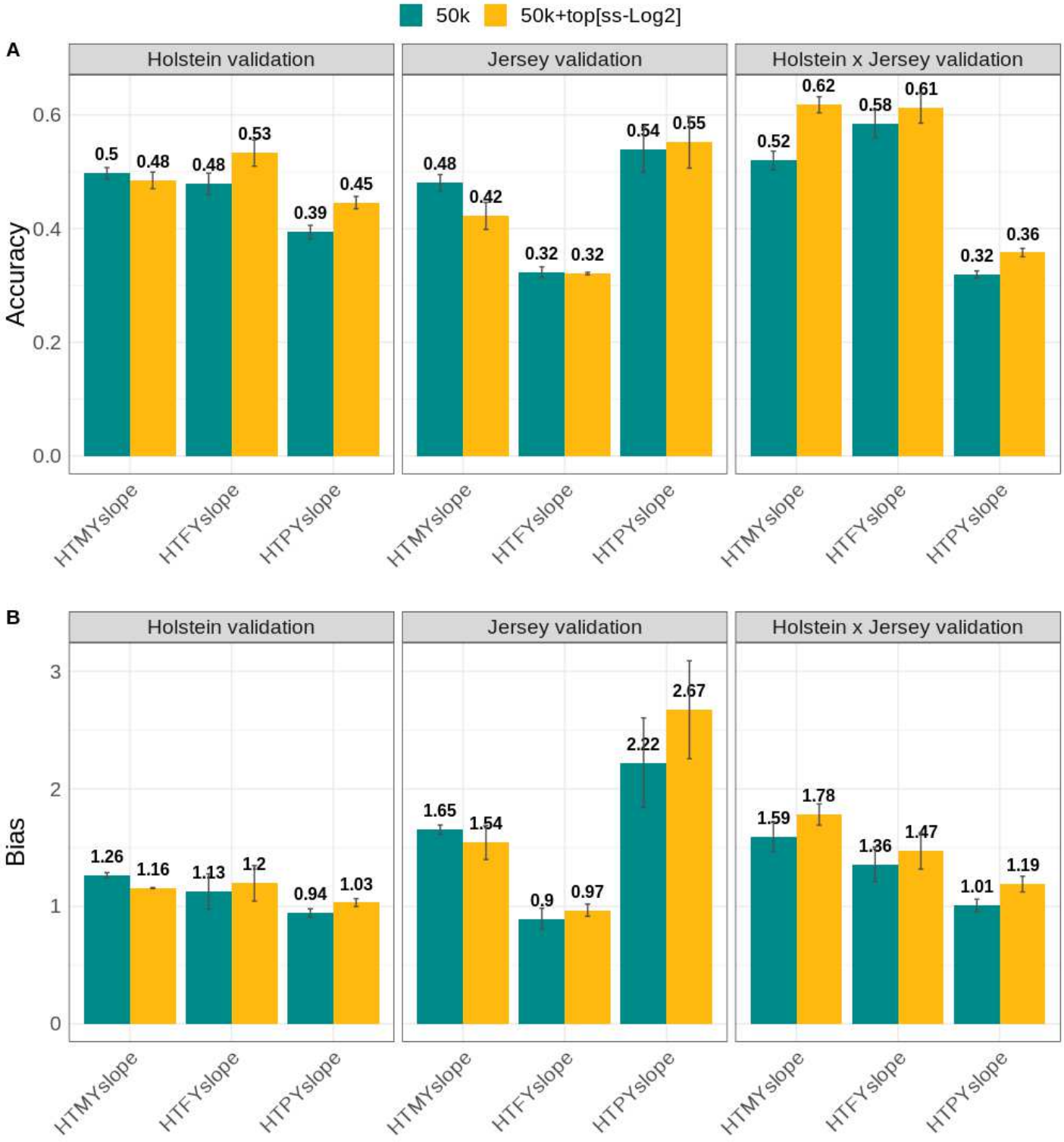

Figure 4

Accuracy and bias of genomic predictions in Holsteins (N = 1,223), Jersey (N = 431) and crossbred (N = 790) cows when using multi-breed reference set (Holstein and Jersey bulls; N = 4,175). The selected ‘top SNPs’ used in the BayesRC were from Holstein cow discovery set (N = 20,623) based on single-trait GWAS cut-off of [-log10(p-value) ≥ 2]. Traits analysed are heat tolerance milk (HTMYslope), fat (HTFYslope), and protein (HTPYslope) yield slopes. Vertical lines represent standard errors calculated from two random validation subsets.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx