University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Kemper, KE;Bowman, PJ;Hayes, BJ;Visscher, PM;Goddard, ME

Title:
A multi-trait Bayesian method for mapping QTL and genomic prediction

Date:
2018-03-24

Citation:
Kemper, K. E., Bowman, P. J., Hayes, B. J., Visscher, P. M. & Goddard, M. E. (2018). A multi-trait Bayesian method for mapping QTL and genomic prediction. Genetics Selection Evolution, 50 (1), https://doi.org/10.1186/s12711-018-0377-y.

Persistent Link:
https://hdl.handle.net/11343/253508

License:
CC BY

Kemper *et al. Genet Sel Evol (2018) 50:10*
https://doi.org/10.1186/s12711-018-0377-y

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

# A multi-trait Bayesian method for mapping QTL and genomic prediction

Kathryn E. Kemper[1,2*], Philip J. Bowman[3,4], Benjamin J. Hayes[3,5,6], Peter M. Visscher[2,7] and Michael E. Goddard[1,3]

## Abstract

**Background:** Genomic prediction and quantitative trait loci (QTL) mapping typically analyze one trait at a time but this may ignore the possibility that one polymorphism affects multiple traits. The aim of this study was to develop a multivariate Bayesian approach that could be used for simultaneously elucidating genetic architecture, QTL mapping, and genomic prediction. Our approach uses information from multiple traits to divide markers into 'unassociated' (no association with any trait) and 'associated' (associated with one or more traits). The effect of associated markers is estimated independently for each trait to avoid the assumption that QTL effects follow a multi-variate normal distribution.

**Results:** Using simulated data, our multivariate method (BayesMV) detected a larger number of true QTL (with a posterior probability > 0.9) and increased the accuracy of genomic prediction compared to an equivalent univariate method (BayesR). With real data, accuracies of genomic prediction in validation sets for milk yield traits with high-density genotypes were approximately equal to those from equivalent single-trait methods. BayesMV tended to select a similar number of single nucleotide polymorphisms (SNPs) per trait for genomic prediction compared to BayesR (i.e. those with non-zero effects), but BayesR selected different sets of SNPs for each trait, whereas BayesMV selected a common set of SNPs across traits. Despite these two dramatically different estimates of genetic architecture (i.e. different SNPs affecting each trait vs. pleiotropic SNPs), both models indicated that 3000 to 4000 SNPs are associated with a trait. The BayesMV approach may be advantageous when the aim is to develop a low-density SNP chip that works well for a number of traits. SNPs for milk yield traits identified by BayesMV and BayesR were also found to be associated with detailed milk composition.

**Conclusions:** The BayesMV method simultaneously estimates the proportion of SNPs that are associated with a combination of traits. When applied to milk production traits, most of the identified SNPs were associated with all three traits (milk, fat and protein yield). BayesMV aims at exploiting pleiotropic QTL and selects a small number of SNPs that could be used to predict multiple traits.

## Background

Understanding the genetics of quantitative or complex traits has been revolutionized by the availability of dense panels of single nucleotide polymorphisms (SNPs) that cover the genome. Data on SNP genotypes combined with phenotypic measurements have been used for three purposes: to study the genetic architecture of quantitative traits, to map regions of the genome that cause variation in these traits (quantitative trait loci or QTL) and to predict the genetic or breeding values of individuals for quantitative traits. Although different statistical methods are commonly used for these three purposes, we have argued that a non-linear Bayesian method that fits all SNPs simultaneously can be used for all three aims [1, 2]. For example, BayesR makes across-breed predictions of breeding values and maps QTL more accurately than genomic best linear unbiased prediction (GBLUP) [1].

QTL often affect more than one trait [3, 4] but most methods that are applied to analyze SNP data use only one trait at a time. Multivariate analyses have been

*Correspondence: k.kemper@uq.edu.au
[2] Institute of Molecular Biosciences, University of Queensland, St. Lucia 4072, Australia
Full list of author information is available at the end of the article

BioMed Central

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

found to increase power to detect and map QTL [5] and increase the accuracy of estimated breeding values (EBV) [6]. For instance, QTL mapping is frequently performed in genome-wide association studies (GWAS) by single-SNP regression, in which the effect of one SNP at a time on the trait is tested. Multiple-trait versions of single-SNP regression have been implemented in various studies e.g. [7–9]. Multi-trait EBV are frequently calculated from pedigree data or from SNP genotypes using BLUP, and sometimes have a higher accuracy compared to single-trait EBV [10, 11]. In dairy cattle, multi-trait genomic prediction models using multi-breed populations (i.e. where a trait measured in different breeds is treated as multiple traits) have been attempted in several studies to account for between-breed differences in QTL effects but with limited success [12–16]. The focus of our paper is the development of a Bayesian multi-trait genomic prediction method for multiple different traits.

Jia and Jannink [17] and Calus and Veerkamp [18] have described non-linear Bayesian methods for the analysis of high-density SNP data on multiple traits. However, these methods assume that the effects of a QTL on different traits are drawn from a multivariate normal distribution with the same correlation for all QTL, which may be incorrect. For instance, an allele of the gene *DGAT1* (*diacylglycerol O-acyltransferase 1*) increases milk yield but decreases milk fat yield in spite of a positive overall genetic correlation between milk and fat yields [19]. Kemper et al. [1] reported numerous similar cases where the pattern of effects of QTL on traits varied, while the overall genetic correlation between milk traits was positive. Some QTL may also affect two traits although there is a weak genetic correlation between the traits. Thus, the assumption of multivariate normality may be too strong.

The aim of our study was to develop a multivariate version of BayesR that uses multiple trait data to decide which SNPs should be included in the model but allows flexibility in the estimation of the effect of the selected SNPs on each of the traits analyzed by estimating the effect of selected SNPs independently for each trait. The hypothesis under investigation is that multi-trait information will improve accuracies of genomic prediction. We illustrate the method by applying it to simulated data for QTL mapping and genomic prediction, and to real data on milk, fat and protein yields from dairy cattle.

## Methods

### Real data

The available dairy cattle dataset had over 16,000 records on Holstein and Jersey cattle from a previous study [1]. The reference population consisted of up to 11,527 Holstein and 4687 Jersey animals, while the validation dataset (used to evaluate the accuracy of genomic predictions) consisted of phenotype records for up to 262 Holstein bulls, 105 Jersey bulls, and 361 Australian Red (bull and cow) cattle (Table 1). Australian red cattle were never included in the reference population, so validation for these animals represents across-breed prediction. Australian red cattle are more closely related to Holstein than to Jersey cattle, as detailed in "Appendix 1".

Phenotypes are three yield traits (fat, milk and protein yield, i.e. FY, MY and PY, respectively) supplied as either daughter-trait deviations (for males) or trait-deviations (for females) from the Australian Dairy Herd Improvement Scheme [1]. Heterogeneous error variances in the phenotypes were accounted for by using a weighted analysis, using the weighting procedure outlined by Garrick et al. [20] and described in detail for this dataset by Kemper et al. [1]. Trait heritabilities and phenotypic and genetic correlations between traits were estimated using the relationship matrix constructed from pedigree data using ASReml [21]. All animals had real or imputed genotypes for 632,002 SNPs from the bovine HD array [1, 22]. Full details of the imputation and quality control procedures for genotypes are described in [22] but they include removal of SNPs with very poor imputation accuracy and SNPs with low minor allele frequency (less than 10 copies in the dataset).

GWAS summary statistics (i.e. allele name, allele effect and standard deviation) were available from a

**Table 1 Number of records in the reference and validation datasets for Holstein, Jersey and Australian Red dairy cattle**

| Breed | Traits | Total records | Reference | | Validation | | |
|---|---|---|---|---|---|---|---|
| | | | Bulls | Cows | YOB | Bulls | Cows |
| Holstein | FY, MY, PY | 11,789 | 3049 | 8478 | 2005 | 262 | – |
| Jersey | FY, MY, PY | 4793 | 770 | 3917 | 2005 | 105 | – |
| Australian Red* | FY, MY, PY | 361 | – | – | – | 114 | 247 |

Year-of-birth (YOB) for the animals included in the validation datasets is also provided

FY = fat yield (kg/lac), MY = milk yield (L/lac); PY = protein yield (kg/lac)

*Australian Red animals were only used for validation and never included in the reference population

previous study [23] which used a subset of Holstein cows (N = 444) from this dataset. These statistics were for detailed milk phenotypes and we used them to verify QTL that were identified using the Bayesian approaches.

### Simulated data

Additive QTL were simulated on real dairy cattle genotypes for a single chromosome to illustrate the impact of the multi-trait method on (1) the accuracy of genomic predictions, (2) the ability to elucidate genetic architecture and (3) the power of QTL mapping. In the simulation, genotypes consisted of 12,745 SNPs from *Bos taurus* autosome 29 (BTA29) for 3049 Holstein animals from the larger reference dataset described above. Simulations involved two traits with 10 QTL per trait, two scenarios with a different number of pleiotropic QTL and 20 replicates for each scenario. In scenario 1, there were no pleiotropic QTL (pQTL = 0) and 20 SNPs were randomly selected as QTL with effects on either trait 1 or trait 2, separately for each replicate. In scenario 2, all QTL had pleiotropic effects (pQTL = 1) and thus, 10 SNPs were randomly selected for each replicate to have effects on both traits. All SNPs chosen as QTL had a minor allele frequency higher than 0.01. Additive QTL effects explained 0.01 of the phenotypic variance ($\sigma_P^2$) and were calculated for each chosen SNP as $\sqrt{0.01\sigma_P^2/2a(1-a)}$, where $a$ is the allele frequency of the SNP, and were randomly allocated to have either positive or negative effects on the trait. Thus, the heritability of the traits was 0.10. The error co-variance between traits was zero.

### Statistical analysis

#### *Construction of traits with independent errors*

Method BayesMV (described later) requires uncorrelated residual errors. Thus, we used a principal component analysis to form linear combinations of the three milk yield traits to create new traits with zero error covariance. A principal component decomposition was conducted on the error correlation matrix (**K**, a $t \times t$ matrix where $t$ = number of traits) from the pedigree-based multivariate analysis in ASReml [21] [see Additional file 1: Table S1]. Then $\mathbf{K} = \mathbf{B}\boldsymbol{\Lambda}\mathbf{B}'$, where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues and **B** is a $t \times t$ matrix of eigenvectors. The linear combinations (**LC**) of traits were constructed by $\mathbf{LC} = \mathbf{B}'\mathbf{T}^{-1}\mathbf{y}$, where **T** is a diagonal matrix of error standard deviations and **y** is the vector of phenotypes. The genetic variance of the linear combinations is $\mathbf{B}'\mathbf{T}^{-1}\mathbf{G}\mathbf{T}^{-1}\mathbf{B}$, where **G** is the genetic variance–covariance matrix. Estimated breeding values for the original traits were constructed for an individual as $\mathbf{TBLC}_{\mathbf{GEBV}}$, where $\mathbf{LC}_{\mathbf{GEBV}}$ is the vector of EBV for the linear combination traits.

#### *Statistical model*

The model fitted to the data for both BayesR and BayesMV had the general form $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Wv} + \mathbf{e}$, where **y** is the vector of phenotypes (i.e. for the linear combination of traits), **b** is a vector of fixed effects, **a** is a vector of polygenic breeding values not explained by the SNPs [distributed as $N(0, \mathbf{A}\sigma_a^2)$, where **A** is the numerator relationship matrix (from pedigree) and $\sigma_a^2$ is the additive genetic variance not explained by the SNP], **v** is a vector of SNP effects assumed normally distributed [$N(0, \sigma_k^2)$, with $k = 1, 2, 3$ or 4: $\sigma_1^2 = 0$, $\sigma_2^2 = 0.0001\sigma_{a^*}^2$, $\sigma_3^2 = 0.001\sigma_{a^*}^2$ and $\sigma_4^2 = 0.01\sigma_{a^*}^2$, where $\sigma_{a^*}^2$ is the additive genetic variance estimated from pedigree], **W** is a matrix of standardized SNP genotypes (as defined in "Appendix 2"), and **e** is a vector of residual errors [distributed as $N(0, \mathbf{R})$, where **R** is the error covariance matrix, $\mathbf{R} = \mathbf{E}\sigma_e^2$ and $\mathbf{E}^{-1}$ is a diagonal matrix of error weights]. We fitted both the univariate (BayesR) and multivariate (BayesMV) models to the linear combinations. A full description of the univariate BayesR used is in [1], while method BayesMV is described in the next section. BayesMV differs from BayesR in that a proportion of the SNPs ($p$) are defined as 'unassociated' and are assumed to have no effect on any trait. BayesR is equivalent to assuming that $p = 0$, so all SNPs are associated and effects are estimated separately for each trait.

#### *BayesMV*

In the BayesMV method, the traits are analyzed simultaneously. It is assumed that the residuals (**e**) and polygenic effects (**a**) are independent between traits and the only connection between traits is the model for the SNP effects (**v**). SNPs are either 'associated' with prior probability $1 - p$ or 'unassociated' with prior probability $p$. If a SNP is unassociated, it has no effect on any trait. If a SNP is associated, the effects on each trait are assumed to be independent and drawn from a mixture of normal distributions, $N(0, \sigma_k^2)$, with $k = 1, 2, 3$or 4: $\sigma_1^2 = 0$, $\sigma_2^2 = 0.0001\sigma_{a^*}^2$, $\sigma_3^2 = 0.001\sigma_{a^*}^2$ and $\sigma_4^2 = 0.01\sigma_{a^*}^2$, where $\sigma_{a^*}^2$ is the additive genetic variance estimated from pedigree [22]. Thus, the probability that the effects of a SNP are drawn from each distribution depends on $p$ (the probability that the SNP is unassociated) and $q_{j,k}$ (the mixing proportion of distribution $k$ for trait $j$ conditional on the SNP being associated). For example, consider two traits and distribution 1 for trait 1 and distribution 2 for trait 2. The probability that the effect of the SNP belongs to distributions $k1$ and $k2$ is equal to $p + (1-p)q_{1,k1}q_{2,k2}$ when $k1 = k2 = 1$ and equal to $(1-p)q_{1,k1}q_{2,k2}$ otherwise. Full details of the model and the Gibbs sampler used to implement the model are in "Appendix 2". Reported effects are posterior means of samples from the

Gibbs sampler with at least 30,000 iterations, with 20,000 iterations discarded as burn-in. Final results are the mean of five replicate chains.

#### *Accuracy and bias of genomic predictions*

The accuracy of the genomic predictions was assessed in the validation population of young Holstein (N = 262), Jersey (N = 105) bulls, and Australian Red animals (N = 361) (Table 1). In simulations, accuracies were calculated as r(TBV, GEBV), where TBV is the true breeding value and GEBV is the breeding value predicted using SNP genotypes. In the real data, the accuracy of genomic predictions was the correlation of predicted breeding values with daughter-deviations (for bulls) or phenotypes (for cows) in the validation dataset, i.e. $r(\hat{\mathbf{y}}, \mathbf{y})$ where $\mathbf{y}$ is a vector of phenotypes for the validation population and $\hat{\mathbf{y}}$ is a vector of GEBV for the corresponding animals in $\mathbf{y}$. The bias of the predictions was assessed as the regression slope $b(\mathbf{y}, \hat{\mathbf{y}})$, where an unbiased (ideal) prediction has a regression slope of 1. Accuracy and bias for Australian Reds were computed as the average results for bulls and cows.

### Association analyses

A multi-trait single-SNP regression association analysis method [4] was used in the simulated data for QTL mapping for comparison to the Bayesian approaches. GWAS association summary statistics from a previous study [23] were used to verify QTL regions identified by using BayesMV and BayesR. This data consisted of the allele effect and standard deviation for concentration of lactose, minerals (calcium, potassium, magnesium, sodium, phosphorus, sulfate and zinc; mg/kg), and proteins (lactoperoxidase, lactoferrin, immunoglobulinG, alpha-lactalbumin, beta-lactoglobulin, kappa-casein, alpha-S1-casein and beta-casein; mg/g) in milk. Briefly, the analysis was for up to 444 cows for which trait records had been corrected for non-genetic effects, such as herd-year-season and stage of lactation, and a mixed linear model was used to detect associations between these traits and 609,563 autosomal SNPs with a minor allele frequency higher than 0.01. The –mlma-loco option using GCTA [24] was used which fits a genomic relationship matrix to account for population structure, where SNPs from the chromosome under test are excluded from the relationship matrix and the tested SNPs are fitted one-at-a-time as fixed effects.

## Results

### Simulated data

#### *Genetic architecture*

In the simulated dataset, there were 20 and 10 QTL with non-zero effects on one or both traits for scenarios 1 and 2, respectively, and all QTL had a variance of 0.01 $\sigma_P^2$. Table 2 shows the posterior number of SNPs from joint distributions for traits 1 and 2, for BayesR and BayesMV (where the joint distribution for BayesR was calculated as the product of the posterior probabilities for each trait). For example, when pQTL = 0, BayesMV estimated that 25 SNPs were only associated with trait 1, 23 were only associated with trait 2 and 41 SNPs were associated with both traits. Considering only trait 1, the number of SNPs estimated to have no effect (i.e. to have an effect sampled from distribution 1) was equal to 12,569 (12,396 + 173) for BayesR and 12,680 (12,643 + 14 + 23) for BayesMV. BayesR estimated similar distributions for both scenarios; i.e. hundreds of SNPs were associated with either trait 1 or trait 2, and two to three SNPs were associated with both traits. BayesMV estimated the true mixing proportions more accurately especially when the same SNPs affected both traits in the simulated data (pQTL = 1). When the same SNPs affected both traits (pQTL = 1), BayesMV estimated that most of the 17 SNPs classified as associated with both traits (i.e. they were sampled from non-zero distributions) and the remaining two SNPs were associated with either trait 1 or trait 2. However, when different SNPs were simulated to affect the two traits, BayesMV estimated an inflated number of SNPs to be associated with both traits (i.e. 41 when no SNPs were simulated to affect both traits). Both BayesR and BayesMV fitted more SNPs in the model than the number of QTL simulated, probably because they fit multiple SNPs for a given QTL, each with a small variance, instead of only 10 SNPs with a large variance.

**Table 2 Posterior mean number of SNPs allocated to each component of the mixture distribution for the two simulated traits when analyzed with BayesR or BayesMV**

| pQTL[a] | Distribution[b] | Simulated[c] | BayesR[d] | BayesMV |
|---|---|---|---|---|
| 0.0 | Unassociated | 12,725 | – | 12,643 |
| | $(\text{trait1})_1_(\text{trait2})_1$ | 0 | 12,396 | 14 |
| | $(\text{trait1})_{2-4}_(\text{trait2})_1$ | 10 | 174 | 25 |
| | $(\text{trait1})_1_(\text{trait2})_{2-4}$ | 10 | 173 | 23 |
| | $(\text{trait1})_{2-4}_(\text{trait2})_{2-4}$ | 0 | 2 | 41 |
| 1.0 | Unassociated | 12,735 | – | 12,728 |
| | $(\text{trait1})_1_(\text{trait2})_1$ | 0 | 12,435 | 0 |
| | $(\text{trait1})_{2-4}_(\text{trait2})_1$ | 0 | 148 | 1 |
| | $(\text{trait1})_1_(\text{trait2})_{2-4}$ | 0 | 159 | 1 |
| | $(\text{trait1})_{2-4}_(\text{trait2})_{2-4}$ | 10 | 3 | 15 |

[a] QTL were independent (no pleiotropic QTL, pQTL = 0) or completely pleiotropic (pQTL = 1.0)

[b] Subscripts indicate distributions 1 to 4, where distributions 1 to 4 explain 0, 0.0001, 0.001 or 0.01 $\sigma_P^2$, respectively

[c] The number of simulated QTL is also provided

[d] Joint probabilities are the product of posterior probabilities (*p* and *q*)

### *QTL mapping*

Figure 1 illustrates the results from the BayesMV, BayesR, and (the multi-trait) single-SNP regression analyses for a single replicate dataset where all QTL were pleiotropic (pQTL = 1). The mean posterior probability (PP) that a SNP had a non-zero effect on any trait was the criterion for mapping QTL. This quantity is estimated directly in BayesMV and was calculated as 1 – (probability of being in the zero distribution for all traits) in BayesR. In the simulated data, the QTL were among the SNPs analyzed, so the ideal outcome is that these SNPs have a high PP and all others have a low PP. As shown in Fig. 1, the QTL were mapped quite accurately by BayesMV, since all the high PP SNPs were either a causative SNP or very close to one. BayesMV tended to have zero or close to zero PP for the remaining SNPs. Compared to BayesMV, BayesR showed more SNPs with a low PP. It is very difficult from the single-SNP regression GWAS to determine the location and possible number of QTL in the simulated dataset. All methods sometimes assigned the highest PP or − log10(P) to non-causal SNPs, demonstrating the influence of linkage disequilibrium (LD) and other nearby QTL on the association statistics in the cattle genotypes.

Results over the 20 replicate datasets are summarized in Figure S1 [see Additional file 2: Figure S1]. In scenarios where all QTL were pleiotropic (pQTL = 1.0), BayesMV had greater power to detect QTL (i.e. where power = number of detected causative SNPs/total number of causative SNPs) when the PP was high (PP > 0.9). This was achieved with approximately the

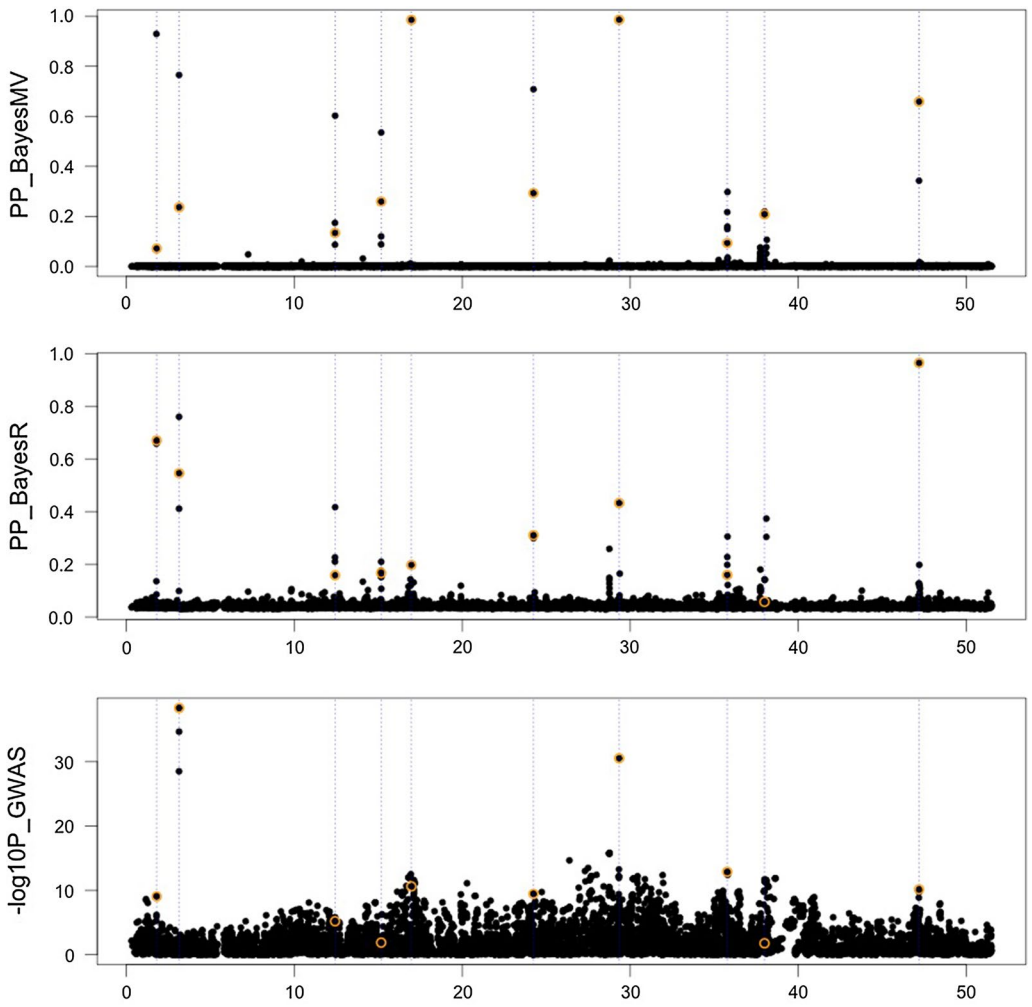



**Fig. 1** QTL mapping in simulated data where all QTL are pleiotropic. The mean posterior probabilities (PP) of SNPs having a non-zero effect for any trait for multivariate (BayesMV, top) and univariate (BayesR, middle) methods with the –log10(*P* value) for the multi-trait single-SNP genome-wide association study (GWAS, bottom) are shown. Results for simulated QTL are highlighted in orange and their position marked with dashed vertical lines

same false-discovery rate (FDR), (i.e. number of detected unassociated SNPs/total number of detected SNPs) as BayesR. In simulations with no pleiotropy, BayesMV had similar power but a higher FDR than BayesR. This was likely caused by SNPs that influenced trait 2 being sampled more frequently from the non-zero distributions for trait 1 (i.e. joint probabilities from BayesMV indicated that there were 41 SNPs that had non-zero effects for both traits but no QTL were simulated to affect both traits).

#### *Genomic prediction*

The accuracy of predictions from the simulated data were high (~ 0.90) when validated in a population similar to the reference dataset (Holstein) but lower when used for across-breed prediction (validated in Jersey, Table 3). Accuracies were much higher than what is typically observed in real data but this is as expected when causative QTL are included in the analyzed set of SNPs, and when the sizes of QTL effects are relatively large (0.01 $\sigma_P^2$). In simulations without pleiotropic QTL (pQTL = 0), BayesR and BayesMV achieved very similar accuracies of prediction, both for the Holstein and the Jersey validation datasets (Table 3). However, when all QTL were pleiotropic (pQTL = 1.0), BayesMV had significantly higher prediction accuracy than BayesR by on average 0.08 for Holstein validation and 0.14 for Jersey validation.

### Real data

#### *Linear decomposition of traits*

Milk yield traits had moderate heritabilities ($h^2$ ~ 0.5) and moderate-high genetic correlations between them (0.5 to 0.8) [see Additional file 1: Table S1]. The eigenvector coefficients for the milk production traits showed that the first linear combination (LC1) was positively correlated to all three yield traits, the second linear combination (LC2) was primarily fat yield corrected for milk and protein yield, and the third linear combination (LC3) was protein yield corrected for milk yield (Table 4). Heritability estimates for the linear combination of traits were moderate to high (0.45 to 0.88, Table 4).

#### *Genetic architecture*

The distribution of SNPs across the four distributions was relatively consistent between BayesR and BayesMV (Table 5), i.e., both methods found that 3000 to 4000 SNPs had non-zero effects for each trait. These findings were also consistent with the analysis of the original milk yield traits [see Additional file 3: Table S2]. The difference between the methods was that BayesMV explained all three traits by the same SNPs. For example, only 4092 associated SNPs were identified in the combined Holstein/Jersey reference set, most of which had effects for all three traits. For instance, joint probabilities identified only one SNP in the associated class that had a zero effect for all three traits, 79 SNPs with effects for only one trait, 949 SNPs with effects for two traits, and 3062 SNPs with effects for all three traits. In contrast, for BayesR, although the number of SNPs with non-zero effects per trait was similar to that observed for BayesMV, almost

**Table 4 Eigenvectors applied to each trait to construct linear combinations (LC1, 2 and 3) with zero error co-variance for the milk production traits**

| | LC1 | LC2 | LC3 |
|---|---|---|---|
| Fat yield | 0.55 | 0.83 | − 0.01 |
| Milk yield | 0.59 | − 0.39 | − 0.70 |
| Protein yield | 0.59 | − 0.39 | 0.71 |
| Heritability ($h^2$)[a] | 0.45 | 0.73 | 0.88 |

[a] Estimated heritability of the linear combinations

**Table 3 Accuracy of genomic predictions for the analysis of the two simulated traits with BayesR and BayesMV for two pleiotropy scenarios**

| Method | Pleiotropy[a] | Trait | Holstein | | Jersey | |
|---|---|---|---|---|---|---|
| | | | Accuracy | SE | Accuracy | SE |
| BayesR | pQTL = 0 | 1 | 0.88 | 0.07 | 0.78 | 0.15 |
| | | 2 | 0.88 | 0.04 | 0.82 | 0.10 |
| | pQTL = 1.0 | 1 | 0.89 | 0.05 | 0.77 | 0.15 |
| | | 2 | 0.89 | 0.04 | 0.80 | 0.14 |
| BayesMV | pQTL = 0 | 1 | 0.89 | 0.06 | 0.80 | 0.15 |
| | | 2 | 0.90 | 0.04 | 0.84 | 0.11 |
| | pQTL = 1.0 | 1 | 0.97 | 0.02 | 0.96 | 0.04 |
| | | 2 | 0.97 | 0.01 | 0.96 | 0.04 |

*SE* standard error (across replicates)

[a] QTL were independent (no pleiotropy, pQTL = 0) or completely pleiotropic (pQTL = 1.0) and accuracies are for validation within-(Holstein) or across (Jersey) breeds

**Table 5 Posterior mean number of SNPs[a] in each distribution for milk production traits from BayesMV or BayesR**

| Reference | Distribution[b,c] | BayesR | BayesMV |
|---|---|---|---|
| Hol_Jer | Unassociated | – | 627,911 |
| | $LC1_1_LC2_1_LC3_1$ | 620,515 | 1 |
| | $LC1_1_LC2_1_LC3_{2-4}$ | 3504 | 11 |
| | $LC1_1_LC2_{2-4}_LC3_1$ | 2994 | 4 |
| | $LC1_{2-4}_LC2_1_LC3_1$ | 4913 | 64 |
| | $LC1_1_LC2_{2-4}_LC3_{2-4}$ | 21 | 47 |
| | $LC1_{2-4}_LC2_1_LC3_{2-4}$ | 29 | 685 |
| | $LC1_{2-4}_LC2_{2-4}_LC3_1$ | 25 | 218 |
| | $LC1_{2-4}_LC2_{2-4}_LC3_{2-4}$ | 1 | 3062 |
| Holstein | Unassociated | – | 628,451 |
| | $LC1_1_LC2_1_LC3_1$ | 621,268 | 0 |
| | $LC1_1_LC2_1_LC3_{2-4}$ | 2817 | 2 |
| | $LC1_1_LC2_{2-4}_LC3_1$ | 3110 | 4 |
| | $LC1_{2-4}_LC2_1_LC3_1$ | 4743 | 12 |
| | $LC1_1_LC2_{2-4}_LC3_{2-4}$ | 17 | 50 |
| | $LC1_{2-4}_LC2_1_LC3_{2-4}$ | 22 | 124 |
| | $LC1_{2-4}_LC2_{2-4}_LC3_1$ | 25 | 234 |
| | $LC1_{2-4}_LC2_{2-4}_LC3_{2-4}$ | 0 | 3124 |
| Jersey | Unassociated | – | 630,779 |
| | $LC1_1_LC2_1_LC3_1$ | 624,314 | 0 |
| | $LC1_1_LC2_1_LC3_{2-4}$ | 1957 | 3 |
| | $LC1_1_LC2_{2-4}_LC3_1$ | 1366 | 3 |
| | $LC1_{2-4}_LC2_1_LC3_1$ | 4335 | 1 |
| | $LC1_1_LC2_{2-4}_LC3_{2-4}$ | 6 | 101 |
| | $LC1_{2-4}_LC2_1_LC3_{2-4}$ | 14 | 28 |
| | $LC1_{2-4}_LC2_{2-4}_LC3_1$ | 10 | 30 |
| | $LC1_{2-4}_LC2_{2-4}_LC3_{2-4}$ | 0 | 1057 |

[a] The posterior mean number of unassociated SNPs from BayesMV is shown with the joint probability of a non-zero effect on one or more traits. Joint probabilities are the product of posterior probabilities (*p* and *q*)

[b] Traits are three linear combinations (LC1, LC2, LC3) of fat, milk and protein yield

[c] Subscripts indicate distributions 1 to 4, each explaining 0, 0.0001, 0.001 or 0.01 of the genetic variance

11,500 SNPs had effects for at least one trait and only one SNP had effects for all three traits. Both models estimated that the Holstein or Jersey reference populations each had fewer associated SNPs (i.e. with non-zero effects) than the combined breed reference population. This is as expected if some QTL segregate in one breed only.

### *QTL mapping*

QTL mapping by BayesMV was assessed in the real data by investigating the top 100 SNPs with the highest PP for inclusion in the model [see Additional file 4: Table S3]. These 100 SNPs were grouped into four types of QTL based on their pattern of effects on milk, fat and protein yield, where the effect of a SNP for a trait was transformed from the LC as $\mathbf{TBLC_{SNP}}$, with $\mathbf{LC_{SNP}}$ being the vector of SNP effects for the linear combinations and **B** the matrix of eigenvectors from Table 4. The largest group of SNPs corresponded to those that had opposite effects on fat yield versus milk volume and protein yield. This group included several previously mapped loci such as *DGAT1* [19], *GPAT4* (*glycerol-3-phosphate acyltransferase 4*) [25] and *MGST1* (*microsomal glutathione S-transferase 1*) [26]. The second largest group of SNPs corresponded to those where an allele increased milk volume and milk solids (fat and protein yields). In some cases, two SNPs that were less than 100 Mbp apart showed a high PP but different patterns of effects, potentially indicating two or more QTL (i.e. BTA3 at about 15.5 Mbp). We chose two regions on chromosomes 11 and 19 for further investigation, where SNPs with a high PP from the BayesMV analysis also overlapped with SNPs associated with either lactoglobulin or potassium concentration [see Additional file 5: Table S4].

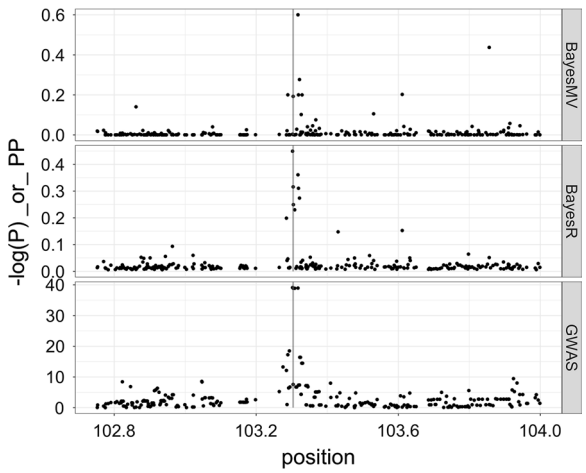



**Fig. 2** Mean posterior probability and associations with beta-lactoglobulin concentration on bovine chromosome 11. The posterior probability (PP) for a SNP being associated with any trait from BayesMV and BayesR and the single-SNP association analysis results are shown. The *PAEP* coding region (formally *LGB*, *lactoglobulin beta*) is highlighted in grey

The mean PP from BayesMV and BayesR on chromosome 11 is shown in Fig. 2. The SNP identified by BayesMV is located downstream of *PAEP* (*progestagen associated endometrial protein*; formally known as *LGB* or *lactoglobulin beta*), while the SNP identified by BayesR is within the coding region of *PAEP*. The figure also shows the GWAS results for beta-lactoglobulin concentration, for which the SNPs identified by both BayesMV (BovineHD1100030073) and BayesR (BovineHD1100030066) were among those that were

most highly associated with beta-lactoglobulin concentration ($P < 1 \times 10^{-39}$; effect ≈ 0.68 mg/g). Although a QTL near *LGB* is known to affect milk volume, fat and protein yields [27], identification of the causal mutation for this QTL has been difficult due to strong LD in the region. The effect of the SNP that was identified by BayesMV had opposite effects on fat yield versus milk and protein yield.

The second QTL investigated was a novel region on chromosome 19, near the *KCNJ2* (*potassium voltage-gated channel subfamily J member 2*) gene [see Additional file 6: Figure S2]. The BovineHD1900017548 SNP was the most significant variant in the region for potassium concentration ($P = 1.21 \times 10^{-7}$; effect = − 62.87 mg/kg) and had a high PP with both BayesR and BayesMV. This SNP had opposite effects on protein yield versus milk and fat yield. KCNJ2 is potassium transporter that tends to transport potassium into (rather than out of) the cell (NCBI GeneID: 3759 [28]).

#### *Genomic prediction*

Genomic prediction accuracies for milk yield traits were moderate to high for breeds that were included in the reference population (~ 0.65 for FY, 0.62 for MY, and 0.58 for PY in Holstein; 0.57 for FY, 0.69 for MY, and 0.71 for PY in Jersey) but relatively low for Australian Reds, which were not included in the reference population (~ 0.27 for FY, 0.18 for MY and 0.09 for PY; Table 6). Results from linear combinations were consistent with previous BayesR analyses of traits in their original form, which were higher than obtained in a previous univariate analysis using GBLUP [see Additional file 7: Table S5]. The BayesMV analysis tended to have similar or slightly higher prediction accuracies than the BayesR analyses when validated in the Holstein set (Table 6) but similar or slightly lower accuracies for the Jersey and the Australian Reds validation sets. This could be due to fewer SNPs having non-zero effects in the BayesMV compared to the BayesR analysis and SNP effects being estimated in a reference population composed mostly of Holstein animals. Estimates of bias were similar to those from a previous analysis of the traits [1] and biases failed to show consistent differences, either when comparing bias of results for the original traits to those for the linear combinations of traits, or when comparing univariate and multivariate analyses of the linear combinations.

## Discussion

Many genomic prediction and QTL mapping methods consider only one trait at a time. In this paper, we present a multivariate method for simultaneous QTL mapping,

**Table 6 Accuracy and bias of genomic predictions for milk production[a] traits using different reference populations and different analysis methods and when validated in Holstein, Jersey or Australian Red animals**

| Analysis method[b] | Reference dataset | Validation dataset | Accuracy[c] | | | Bias | | |
|---|---|---|---|---|---|---|---|---|
| | | | FY | MY | PY | FY | MY | PY |
| BayesR[d] | Holstein | Holstein | 0.63 | 0.62 | 0.58 | 1.22 | 0.89 | 1.02 |
| BayesR_LC | Holstein | Holstein | 0.65 | 0.62 | 0.57 | 1.17 | 0.91 | 0.99 |
| BayesMV | Holstein | Holstein | 0.65 | 0.63 | 0.59 | 1.21 | 0.89 | 1.03 |
| BayesR[d] | Hol_Jer | Holstein | 0.65 | 0.63 | 0.58 | 1.25 | 0.89 | 0.99 |
| BayesR_LC | Hol_Jer | Holstein | 0.65 | 0.62 | 0.58 | 1.14 | 0.90 | 0.97 |
| BayesMV | Hol_Jer | Holstein | 0.66 | 0.63 | 0.58 | 1.17 | 0.87 | 0.97 |
| BayesR[d] | Jersey | Jersey | 0.56 | 0.70 | 0.72 | 0.89 | 0.98 | 1.24 |
| BayesR_LC | Jersey | Jersey | 0.57 | 0.70 | 0.72 | 0.70 | 1.05 | 1.17 |
| BayesMV | Jersey | Jersey | 0.55 | 0.70 | 0.71 | 0.81 | 1.00 | 1.11 |
| BayesR[d] | Hol_Jer | Jersey | 0.56 | 0.69 | 0.71 | 0.93 | 0.95 | 1.18 |
| BayesR_LC | Hol_Jer | Jersey | 0.58 | 0.69 | 0.73 | 0.92 | 1.00 | 1.20 |
| BayesMV | Hol_Jer | Jersey | 0.55 | 0.66 | 0.69 | 0.92 | 0.96 | 1.15 |
| BayesR[d] | Hol_Jer | Aust Red | 0.26 | 0.22 | 0.10 | 0.89 | 0.56 | 0.38 |
| BayesR_LC | Hol_Jer | Aust Red | 0.28 | 0.20 | 0.12 | 0.87 | 0.53 | 0.41 |
| BayesMV | Hol_Jer | Aust Red | 0.26 | 0.14 | 0.07 | 0.75 | 0.34 | 0.25 |

[a] Milk production traits were fat yield (FY), milk yield (MY) and protein yield (PY)

[b] Methods were either BayesR on raw phenotypes (BayesR), linear combinations of traits analyzed with univariate BayesR (BayesR_LC) or the multivariate BayesMV method

[c] Standard errors are approximately 0.062 for Holstein, 0.098 for Jersey and 0.074 for Australian Red predictions

[d] Univariate results from Kemper et al. [1]

analysis of genetic architecture, and genomic prediction, i.e. BayesMV (a multivariate form of BayesR). BayesMV uses information on multiple traits for the selection of SNPs to be included in the model and within-trait information for estimation of the effects of the selected SNPs. It assumes that SNPs fall into one of two classes: either they have no effect for any trait or they have an effect for one or more traits.

The results using simulated data showed that BayesMV can have three advantages over BayesR: BayesMV correctly identifies QTL that affect multiple traits, it maps the QTL more precisely, and it predicts breeding values with greater accuracy. However, when no pleiotropic SNPs were simulated, BayesMV still identified some SNPs with effects for both simulated traits. High false-discovery rates when no pleiotropic QTL are present seems a common problem for multi-trait methods that assume pleiotropic effects [9]. The size and number of QTL that affect traits also influences the performance of multi-trait methods. For example, Jia and Jannink [17] used simulation to show that non-linear multivariate methods can outperform multivariate GBLUP in terms of genomic prediction accuracy when QTL of large effect segregate for traits. Furthermore, Chen et al. [12] showed that improvements in the accuracy of genomic predictions using a multivariate non-linear Bayesian approach were modest (compared to univariate methods) when the traits were affected by many segregating QTL compared to traits with few QTL of large effect. Experimental power to detect QTL (i.e. sample size) also influences the conclusions drawn, with non-linear methods tending to have higher accuracy for genomic predictions compared to GBLUP as sample size increases [1]. Thus, our simulation results should be interpreted with caution since they depend on how the simulations are constructed. However, they do demonstrate that BayesMV can outperform single-trait methods when the genetic architecture of the analyzed traits matches the assumptions that underlie the BayesMV model.

For the real milk yield data from dairy cattle, the advantage of BayesMV over BayesR was not as clear. For example, BayesMV showed limited advantage over BayesR in accuracy of genomic prediction, which is not entirely unexpected. With pedigree-based BLUP, multi-trait EBV are only slightly more accurate than single-trait EBV when all traits are measured on all animals and the traits have similar heritability [29]. Using genomic prediction for psychiatric disorders, Maier et al. [30] observed a 2 to 3% increase in predictive ability for schizophrenia, bipolar disorder, and major depressive disorder when using multivariate GBLUP compared to the univariate models. These data included a large number of missing records (i.e. each individual was recorded for only one trait), traits with moderate and similar heritabilities (~ 0.25), and moderate genetic correlations between traits (~ 0.4 to 0.6). These results agree with several simulations that used GBLUP multivariate genomic predictions [17, 18]. Thus, for genomic prediction, the advantage of multivariate over univariate GBLUP depends on the genetic correlation between traits and the number of (new) records contributed by the trait(s) added to the analysis. Similar conclusions have been drawn for multi-trait versus univariate GWAS models [9] and for prediction accuracies of non-linear multivariate versus univariate models [17]. That is, multivariate models increase the accuracy of genomic predictions for low heritability traits that have a strong genetic correlation with a trait that has moderate to high heritability and when additional records measured on a trait with high heritability are added to the analysis.

Studies that implement multi-trait genomic prediction methods are often motivated by multi-breed prediction problems and use the flexibility that is inherent in these models to share information across breeds [12–16]. For example, BayesRS uses location-specific priors to share information across breeds about the location of QTL regions [14], while other implementations accumulate evidence for QTL across multiple breeds and estimate SNP effects within breed [12, 13]. Thus, information that is shared in both the multi-breed approach and our approach impacts the SNP selection step, rather than the estimation step, of the analysis. Our method could easily be applied to multi-breed genomic prediction problems (with the simplification that there is no error in the covariance between traits).

Single-trait methods, such as BayesR, often identify SNPs that are close to each other but where each SNP appears to be associated with a different trait. The question then arises whether there is one QTL that affects both traits (pleiotropy) or two linked QTL that each affect one trait. The BayesMV method attempts to solve this question but its ability to do so is limited by the information contained in the data. BayesMV makes no assumptions about presence of pleiotropic (one QTL affecting more than one trait) versus linked QTL (multiple QTL each affecting one trait). Any particular SNP can have effects for none, some, or all traits. In the simulated data, BayesMV estimates the traits that each SNP was associated with moderately well but not perfectly, as shown by the results. The high degree of LD in dairy cattle populations may make it difficult to distinguish between pleiotropic QTL and multiple non-pleitropic QTL that are in high LD and each associated with one trait. Fitting all SNPs simultaneously partly accounts for LD but it is impossible to distinguish between pleiotropy and linkage if the linked QTL are in perfect LD. Thus,

although our method is motivated by pleiotropic QTL, we cannot distinguish pleiotropy from tightly linked loci. The distinction, between pleiotropy and linkage, is important for QTL discovery but is inconsequential for genomic prediction.

In the real data, the five independent chains for the BayesMV implementation all showed that most SNPs either have no effect or have effects for all three traits. These results are markedly different from the single-trait BayesR results, where less than 100 SNPs had effects for more than one trait (Table 5). In spite of these apparent differences in estimated architectures, the accuracy of prediction was similar between the multi-trait and single-trait methods (Table 6). We speculate that the high LD in cattle, coupled with the traits being highly polygenic and having many more predictors than records ($p \gg n$) results in many possible solutions for prediction that yield similar accuracies. However, our results do show that one of these possible solutions includes a situation where most selected SNPs have effects for both milk volume and composition. Further information on pleiotropy versus linkage can be gained from the pattern of effects of each SNP across the three traits. For instance, the region identified on chromosome 3 appeared to harbor two close QTL, one at 15.4 Mbp affecting protein yield and one at 15.6 Mbp affecting milk yield (see Additional file 4: Table S3). In contrast, the SNPs on chromosome 14 at 1 to 2 Mbp, all had the same pattern of effects for the three traits and may all be tracking *DGAT1*. In this case, BayesMV may fit multiple SNPs for one QTL because the causal mutation is not in the data and because of the uncertainty caused by limited sample size.

We analyzed uncorrelated linear combinations of the three traits to reduce computing time and to minimize pleiotropy. Since milk, fat and protein yields are highly genetically correlated, the presence of pleiotropic QTL is highly plausible. Therefore, an analysis based on the three original traits was certain to find extensive pleiotropy. The transformation that was applied (based on the residual covariance matrix) does not guarantee zero genetic correlations but in the case of milk, fat and protein yields, the genetic correlations of the transformed traits were assumed low. However, a general procedure would be a canonical transformation, e.g. [31], where variables are both environmentally and genetically uncorrelated.

A consistent finding for both the BayesMV and the univariate BayesR analyses was that 3000 to 4000 SNPs have non-zero effects for milk yield traits. The BayesMV analysis may imply that most of the QTL tracked by these SNPs affect all three traits. The GWAS of detailed milk composition supported this conclusion by showing that, in some cases at least, the same SNPs were associated with additional milk composition traits, such as the concentration in milk of beta-lactoglobulin and potassium.

The major difference that we observed between the BayesMV and BayesR methods was that BayesMV identifies SNPs with effects for all traits in the analysis. Using the linear combinations of traits, BayesMV identified a subset of SNPs that adequately explained variation in multiple traits. This has potential practical value because it means that we could identify a limited number of SNPs that could be genotyped rather than imputed and these SNPs could be used for multiple-trait EBV calculation. However, LD between the causal mutations and SNPs may differ between breeds, which may limit the realized advantage of BayesMV in multi-breed data (compared to univariate methods), particularly when relying on high-density SNPs. The application of BayesMV to sequence data could overcome this limitation and should increase power to identify the causal variants over univariate methods. If many unrelated traits are combined in an analysis, it is possible that a very large number of sequence variants (~ 100 K [32]) will have estimated effects for at least one trait, but the number of associated SNPs should still be much smaller than the total number of polymorphisms in the genome. This would especially be the case if the hypothesis of universal pleiotropy holds, i.e. a mutation at any locus has the potential to affect many (or most) traits [33, 34]. Simultaneously analyzing traits within a physiological domain, as we have done here, is a practical first-step to assessing and using pleiotropy.

## Conclusions

We have implemented a multivariate version of the BayesR methodology that is designed to exploit pleiotropic effects of causal loci to improve mapping ability and, in turn, improve accuracy of genomic predictions. A key feature of our method is that across-trait information is used in the selection of SNPs but effects of SNPs are estimated independently for each trait. Our model performed well in simulated datasets where causal mutations were included in the analyzed SNPs and the QTL had relatively large effects (1% of phenotypic variance). In real data, the multivariate model identified most selected SNPs to be associated with all three milk yield traits (fat, milk and protein yield) but we found little evidence to support our hypothesis that multi-trait information would improve genomic prediction accuracies in milk yield traits for dairy cattle. An advantage of the new method is that it selects a small subset of SNPs that could be used for genomic prediction for multiple traits.

## Additional files

**Additional file 1: Table S1.** Phenotypic and genetic correlation between milk yield traits, with trait heritability from the pedigree-based multi-trait model.

**Additional file 2: Figure S1.** Mean power and false-discovery rate for QTL discovery in simulated data for a single trait.

**Additional file 3: Table S2.** Posterior mean number of SNPs in each distribution [0, 0.0001, 0.001 or 0.01 of the pedigree estimated genetic variance], data from Kemper et al. [1].

**Additional file 4: Table S3.** Top 100 SNPs with the highest mean posterior probability (PP) for inclusion in the model from the Holstein/Jersey reference population using BayesMV [39–41].

**Additional file 5: Table S4.** This file contains highly significant SNPs ($P < 1 \times 10^{-6}$) from the mixed model analysis of the milk minerals and proteins.

**Additional file 6: Figure S2.** Mean posterior probability for BayesMV and BayesR, and the $-\log_{10}(P)$ association test statistic between SNP and potassium concentration on bovine chromosome 19 near *KCNJ2*.

**Additional file 7: Table S5.** Genomic prediction accuracy and bias from the univariate GBLUP model, data from Kemper et al. [1].

**Additional file 8: Figure S3.** Genetic relationship between nine dairy and beef cattle breeds.

#### Authors' contributions

KEK performed the analysis and wrote the paper; PJB implemented the c++ multi-threaded version of BayesMV; BJH and PMV contributed to formulation of the methods and revised the manuscript; MEG supervised the analysis and wrote the paper. All authors read and approved the final manuscript.

#### Author details

[1] Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville 3052, Australia. [2] Institute of Molecular Biosciences, University of Queensland, St. Lucia 4072, Australia. [3] Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, VIC 3083, Australia. [4] School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia. [5] Queensland Agriculture and Food Alliance, University of Queensland, St. Lucia 4072, Australia. [6] Dairy Futures Co-operative Research Centre, Bundoora 3083, Australia. [7] Queensland Brain Institute, University of Queensland, St. Lucia 4072, Australia.

#### Acknowledgements

We thank the Australian Dairy Herd Improvement Scheme and the many dairy farmers who participated in sample collections for their assistance and on-going support.

#### Competing interests

The authors declare that they have no competing interests.

#### Availability of data and materials

Genotype data and trait data for bulls used in this study were kindly provided by the Australian Dairy Herd Improvement Scheme, and are the property of the dairy farmers of Australia and, thus, are not publically available. The milk protein and component data are the property of the dairy farmers that participated in the project, and are also not publically available. However, summary statistics for SNPs with a high posterior probability or significantly associated ($P < 1 \times 10^{-6}$) with milk component traits are available in Additional file 4: Table S3. The BayesMV software is available for non-commercial purposes upon request.

#### Consent for publication

Not applicable.

#### Ethics approval and consent to participate

All data used in this manuscript were previously published and, as such, no new animal ethics approvals were required. Permission to access the trait records and DNA samples was given by the Australian Dairy Herd Improvement Scheme.

#### Funding

This research was supported under the Australian Research Council's Discovery Projects funding scheme (Project DP1093502). The views expressed herein are those of the authors and are not necessarily those of the Australian Research Council. We thank the Agriculture Victoria (and its predecessors) for support and the provision of data for this project.

## Appendix 1

Wright's fixation index ($F_{ST}$) and a per-breed measure of inbreeding [$2F_{ST}/(1 + F_{ST})$] were calculated [35] for all pair-wise combinations of the three dairy breeds (Holstein, Jersey, Australian Reds) and six additional beef breeds (Angus, Charolais, Hereford, Limousin, Murray Grey and Shorthorn, from [36]) using 610,123 autosomal SNPs. Beef breeds were included to provide additional context for the genetic relationships between the dairy breeds. The 'ape' package in R [37, 38] was used to construct an unrooted tree of the relationships between the breeds using the neighbor-joining method. The resultant tree [see Additional file 8: Figure S3] indicates that Jerseys are highly differentiated from all other breeds (e.g. $F_{ST}$: Holstein-Jersey = 0.08; $F_{ST}$: AustRed-Jersey = 0.072), including the beef breeds, and that Australian Reds are closely related to Holsteins ($F_{ST}$: AustRed-Holstein = 0.033).

## Appendix 2

The model fitted to each trait is:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Wv} + \mathbf{e},$$

where

$\mathbf{y}$ is a vector of $n$ phenotypes for cows or bulls,

$\mathbf{b}$ is a vector containing the overall mean, breed and sex effects,

$\mathbf{a}$ is a vector of $u$ polygenic breeding values, distributed as $N(0, \mathbf{A}\sigma_a^2)$,

$\mathbf{v}$ is a vector of $m$ SNP effects,

$\mathbf{e}$ is a vector of $n$ residual errors, distributed as $N(0, \mathbf{R})$,

$\mathbf{X}$ is a design matrix allocating phenotypes to mean and sex effects,

$\mathbf{Z}$ is a design matrix allocating phenotypes to polygenic breeding values ($\mathbf{Z} = n$ by $u$ matrix),

$\mathbf{W}$ is a standardized genotype matrix ($\mathbf{W} = n$ by $m$ matrix), where $w$ is the first column of $\mathbf{W}$ and $w = \left(W_l^* - \bar{W}_l^*\right)/\sigma_l, \sigma_l^2 = 2c(1-c), c = \bar{W}_l^*/2$ and $W_l^*$ is the first column of a $n \times m$ matrix of genotype calls,

$\mathbf{A}$ is a numerator relationship matrix,

$\mathbf{R}$ is an error covariance matrix, where $\mathbf{R} = \mathbf{E}\sigma_e^2$ and $\mathbf{E}^{-1}$ is a $n$ by $n$ diagonal matrix of error weights.

Calculation of the weights follows [1, 20], we used $\frac{d(1-r^2)}{4-h^2}$ for bulls and $\frac{t(1-h^2)}{1+(t-1)r-th^2}$ for the cows (where $t$ is the number of records per cow, $r$ is the trait repeatability, $d$ is the number of daughters per sire and $h^2$ is the heritability of the trait).

$\sigma_a^2$ is the additive genetic variance not explained by the SNPs.

$\sigma_e^2$ is the error variance.

It is assumed that the residuals (**e**) and polygenic effects (**a**) are independent between traits. The only connection between traits is the model for the SNP effects (**v**) as explained below.

#### Priors

The mean, breed and sex effects (**b**), and polygenic effects (**a**) were assigned an uninformative normal prior distribution. The prior for the variance parameters ($\sigma_a^2$ and $\sigma_e^2$) were scaled inverse Chi squared distribution with $n-2$ degrees of freedom.

BayesMV introduces (compared to the univariate equivalent BayesR) a parameter $p$, which is the proportion of unassociated SNPs (i.e. SNPs with zero effect on all traits). For SNPs in the associated class $(1-p)$, the prior distribution for the SNP effects (**v**) follows Erbe et al. [22] and assumes SNP effects are from a mixture of four zero mean normal distributions with 0, 0.0001 $\sigma^2_{\boldsymbol{a}*}$, 0.001 $\sigma^2_{\boldsymbol{a}*}$ or 0.01 $\sigma^2_{\boldsymbol{a}*}$ variance (where $\sigma^2_{\boldsymbol{a}*}$ is the genetic variance, as estimated from pedigree analysis) and mixing proportions $q_{k,j}$ for distribution $k$ from trait $j$. The genetic variance was assumed known and determined prior to the analysis, i.e. see Table 4 for heritability estimates. The mixing proportions ($p$ and $q$) follow Dirichlet distributions with one pseudo count in each category. Note that the distribution from which the effect of an associated SNP is drawn for one trait is independent of that for other traits.

#### Gibbs sampling

Note that in the following, the current estimates of the parameters in the Gibbs sampler (e.g. $\tilde{b}$) are distinguished from the final estimates (e.g. $\hat{b}$) using superscripts.

For each trait,

1. Sample $\sigma_e^2$ from a scaled inverse Chi squared distribution with mean equal to $\tilde{e}'E^{-1}\tilde{e}$ and $n-2$ degrees of freedom, where $\tilde{e} = y - X\tilde{b} - Z\tilde{a} - W\tilde{\nu}$, and $\tilde{b}$, $\tilde{a}$ and $\tilde{\nu}$ are the current estimates of the parameters.
2. Sample estimates for the mean and sex effects from a normal distribution with mean $\left[X'R^{-1}X\right]^{-1}X'R^{-1}y^*$ (where $y^*$ is the phenotype $y$ corrected for the current estimates of all other terms in the model) and variance $\left[X'R^{-1}X\right]^{-1}$.
3. Sample polygenic effects for animal $i$ from a normal distribution with mean $\left[Z_i'R_{ii}^{-1}Z_i + A_{ii}^{-1}\sigma_a^{-2}\right]^{-1}Z_i'R_{ii}^{-1}y^*$ (where $Z_i$ is the row corresponding to animal $i$ in **Z**, and $A_{ii}^{-1}$ and $R_{ii}^{-1}$ are the $i$th diagonal elements of $\mathbf{A}^{-1}$ and $\mathbf{R}^{-1}$, respectively) and variance $\left[Z_i'R_{ii}^{-1}Z_i + A_{ii}^{-1}\sigma_a^{-2}\right]^{-1}$.
4. Sample the polygenic variance from a scaled inverse Chi squared distribution with mean $\tilde{a}'A^{-1}\tilde{a}$ and $n-2$ degrees of freedom.

Combining the data for all traits,

5. Sample the effect of each SNP as follows:
   (a) Sample the SNPs as unassociated with probability $\frac{p\prod_J L_{j,1}}{p\prod_J L_{j,1}+(1-p)\prod_J\left(\sum_K L_{j,k}q_{j,k}\right)}$, where $L_{j,k}$ is the likelihood of the SNP being sampled from distribution $k$ for trait $j$, $K$ is the number of distributions (i.e. $K=4$), $J$ is the number of traits, $p$ and $q_{j,k}$ are the mixing proportions and the first distribution for each trait has zero variance. The log-likelihood is calculated as $\ln\left(L_{k,j}\right) = -0.5\left[\ln\left(1+\mathbf{w}'R^{-1}\mathbf{w}\sigma^2_{k,j}\right) + y_j^{*\prime}R^{-1}y_j^* - y_j^{*}R^{-1}\mathbf{w}\nu_{k,j}\right]$ where $\sigma^2_{k,j}$ is the variance of distribution $k$ for trait $j$, $y_j^*$ is the phenotype for trait $j$ corrected for all other current effects in the model, **w** is a vector of genotypes for the current SNP and $\nu_{k,j} = \left[\mathbf{w}'\mathbf{R}^{-1}\mathbf{w} + \sigma_{k,j}^{-2}\right]^{-1}\boldsymbol{w}'\mathbf{R}^{-1}y_j^*$.
   (b) If the SNP is sampled as unassociated, the SNP effect for all traits is set to zero and the algorithm moves onto the next SNP.
   (c) If the SNP is sampled as associated, the distribution $k$ is sampled independently for each trait $j$ with probability $\frac{L_{j,k}q_{j,k}}{\sum_{J,K}L_{j,k}q_{j,k}}$.
   (d) Using the sampled distributions, the effect of each SNP is sampled for each trait from $N(\nu_{k,j'}, \left[\mathbf{w}'\mathbf{R}^{-1}\mathbf{w} + \sigma^2_{k,j}\right]^{-1})$.
6. Update $p$ by sampling from a Dirichlet distribution given by $p \sim Dir(\alpha + \beta_0)$, where $\alpha' = (1,1)$ is the prior and $\beta_0$ is the current number of markers in the associated and unassociated class.

Update $q_{j,k}$ for each trait by sampling from a Dirichlet distribution given by $q_{j,k} \sim \mathrm{Dir}\left(\gamma + \varepsilon_{k,j}\right)$, where $\gamma' = (1,1,1,1)$ is the prior and $\varepsilon_{k,j}$ is the current number of markers with effects sampled from distribution $k$ for trait $j$.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 12 April 2017 Accepted: 5 February 2018
Published online: 24 March 2018

### References

1. Kemper KE, Reich CM, Bowman PJ, van der Jagt CJ, Chamberlain AC, Mason BA, et al. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy for across-breed genomic predictions. Genet Sel Evol. 2015;47:29.
2. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 2015;11:e1004969.
3. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet. 2013;14:483–95.
4. Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K, et al. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in Beef cattle. PLoS Genet. 2014;10:e1004198.
5. Knott SA, Haley CS. Multitrait least squares for quantitative trait loci detection. Genetics. 2000;156:899–911.
6. Henderson CR, Quaas RL. Multiple trait evaluation using relatives' records. J Anim Sci. 1976;43:1188–97.
7. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014;11:407–9.
8. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FCF, Elliott P, Jarvelin MR, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS One. 2012;7:e34861.
9. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018;50:229–37.
10. Christensen OF, Madsen P, Nielsen B, Ostersen T, Su G. Single-step methods for genomic evaluation in pigs. Animal. 2012;6:1565–71.
11. Wolc A, Zhao HH, Arango J, Settar P, Fulton JE, O'Sullivan NP, et al. Response and inbreeding from a genomic selection experiment in layer chickens. Genet Sel Evol. 2015;47:59.
12. Chen L, Li C, Miller S, Schenkel F. Multi-population genomic prediction using a multi-task Bayesian learning model. BMC Genet. 2014;15:53.
13. Calus MP, Goddard ME, Wientjes Y, Bowman PJ, Hayes BJ. Multibreed genomic prediction using multitrait GREML and multitask Bayesian variable selection. J Dairy Sci. 2018. https://doi.org/10.3168/jds.2017-13366.
14. Brøndum RF, Su G, Lund MS, Bowman PJ, Goddard ME, Hayes BJ. Genome position specific priors for genomic prediction. BMC Genomics. 2012;13:543.
15. Haile-Mariam M, Pryce JE, Schrooten C, Hayes BJ. Including overseas performance information in genomic evaluations of Australian dairy cattle. J Dairy Sci. 2015;98:3443–59.
16. Karoui S, Carabano MJ, Diaz C, Legarra A. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet Sel Evol. 2012;44:39.
17. Jia Y, Jannink JL. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics. 2012;192:1513–22.
18. Calus MP, Veerkamp RF. Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol. 2011;43:26.
19. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, et al. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. Genome Res. 2002;12:222–31.
20. Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol. 2009;41:55.
21. Gilmour AR, Gogel BJ, Cullis BR, Thompson R. ASReml user guide release 2.0. Hemel Hempstead: VSN International Ltd; 2006.
22. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95:4114–29.
23. Kemper KE, Littlejohn MD, Lopdell T, Hayes BJ, Bennett LE, Williams RP, et al. Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. BMC Genomics. 2016;17:858.
24. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.
25. Littlejohn MD, Tiplady K, Lopdell T, Law TA, Scott A, Harland C, et al. Expression variants of the lipogenic *AGPAT6* gene affect diverse milk composition phenotypes in *Bos taurus*. PLoS One. 2014;9:e85757.
26. Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T, et al. Sequence-based association analysis reveals an *MGST1* eQTL with pleiotropic effects on bovine milk composition. Sci Rep. 2016;6:25376.
27. Ng-Kwai-Hang KF. A review of the relationship between milk protein polymorphism and milk composition/milk production. In: Proceedings of the international dairy federation seminar "Milk Protein Polymorphism II": 25–27 February 1997; Palmerston North; 1997. P. 22–37.
28. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. Nucleic Acids Res. 2015;43:D36–42.
29. Schaeffer LR. Sire and cow evaluation under multiple trait models. J Dairy Sci. 1984;67:1567–80.
30. Maier R, Moser G, Chen GB, Ripke S, Cross-Disorder Working Group of the Psychiatric Genomics Consortium, Coryell W, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am J Hum Genet. 2015;96:283–9.
31. Ducrocq V, Chapuis H. Generalizing the use of the canonical transformation for the solution of multivariate mixed model equations. Genet Sel Evol. 1997;29:205–24.
32. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169:1177–86.
33. Wright S. Evolution and the genetics of populations, vol. 1. Chicago: University of Chicago Press; 1968.
34. Stearns FW. One hundred years of pleiotropy: a retrospective. Genetics. 2010;186:767–73.
35. Kemper KE, Saxton SJ, Bolormaa S, Hayes BJ, Goddard ME. Selection for complex traits leaves little or no classic signatures of selection. BMC Genomics. 2014;15:246.
36. Bolormaa S, Pryce JE, Kemper KE, Savin K, Hayes BJ, Barendse W, et al. Accuracy of prediction of genomic breeding values for residual feed intake, carcass and meat quality traits in *Bos taurus*, *Bos indicus* and composite beef cattle. J Anim Sci. 2013;91:3088–104.
37. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90.
38. R: A language and environment for statistical computing. http://www.R-project.org. Accessed 1 Jan 2014.
39. Kemper KE, Hayes BJ, Daetwyler HD, Goddard ME. How old are QTL and how widely do they segregate? J Anim Breed Genet. 2015;132:121–34.
40. Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee JH, et al. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res. 2005;15:936–44.
41. Blott S, Kim JJ, Moisio S, Schmidt-Kuntzel A, Cornet A, Berzi P, et al. Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics. 2003;163:253–66.