genes

MDPI

*Article*

# An Integrative Genomic Prediction Approach for Predicting Buffalo Milk Traits by Incorporating Related Cattle QTLs

**Xingjie Hao [1,*,†], Aixin Liang [2,†], Graham Plastow [3], Chunyan Zhang [3], Zhiquan Wang [3], Jiajia Liu [2], Angela Salzano [4], Bianca Gasparrini [4], Giuseppe Campanile [4], Shujun Zhang [2] and Liguo Yang [2,*]**

1 Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
2 Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Huazhong Agricultural University, Wuhan 430070, China
3 Livestock Gentec Center, Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB T6G 2C8, Canada
4 Department of Veterinary Medicine and Animal Productions, University of Naples "Federico II", 80137 Naples, Italy
* Correspondence: xingjie@hust.edu.cn (X.H.); ylg@mail.hzau.edu.cn (L.Y.)
† These authors contributed equally to this work.

**Abstract:** Background: The 90K Axiom Buffalo SNP Array is expected to improve and speed up various genomic analyses for the buffalo (*Bubalus bubalis*). Genomic prediction is an effective approach in animal breeding to improve selection and reduce costs. As buffalo genome research is lagging behind that of the cow and production records are also limited, genomic prediction performance will be relatively poor. To improve the genomic prediction in buffalo, we introduced a new approach (pGBLUP) for genomic prediction of six buffalo milk traits by incorporating QTL information from the cattle milk traits in order to help improve the prediction performance for buffalo. Results: In simulations, the pGBLUP could outperform BayesR and the GBLUP if the prior biological information (i.e., the known causal loci) was appropriate; otherwise, it performed slightly worse than BayesR and equal to or better than the GBLUP. In real data, the heritability of the buffalo genomic region corresponding to the cattle milk trait QTLs was enriched (fold of enrichment > 1) in four buffalo milk traits (FY270, MY270, PY270, and PM) when the EBV was used as the response variable. The DEBV as the response variable yielded more reliable genomic predictions than the traditional EBV, as has been shown by previous research. The performance of the three approaches (GBLUP, BayesR, and pGBLUP) did not vary greatly in this study, probably due to the limited sample size, incomplete prior biological information, and less artificial selection in buffalo. Conclusions: To our knowledge, this study is the first to apply genomic prediction to buffalo by incorporating prior biological information. The genomic prediction of buffalo traits can be further improved with a larger sample size, higher-density SNP chips, and more precise prior biological information.

**Keywords:** buffalo; pGBLUP; genomic prediction; linear mixed model; enrichment; prior biological information



**Citation:** Hao, X.; Liang, A.; Plastow, G.; Zhang, C.; Wang, Z.; Liu, J.; Salzano, A.; Gasparrini, B.; Campanile, G.; Zhang, S.; et al. An Integrative Genomic Prediction Approach for Predicting Buffalo Milk Traits by Incorporating Related Cattle QTLs. *Genes* **2022**, *13*, 1430. https://doi.org/10.3390/genes13081430

Academic Editor: Xiaolei Liu

Received: 12 July 2022
Accepted: 9 August 2022
Published: 11 August 2022



## 1. Introduction

Genomic prediction is becoming increasingly important for animal and plant breeding programs because of its effectiveness in improving selection and reducing costs [1–5]. The application of genomic prediction to humans has also attracted substantial research interest in terms of human disease prevention and personalized medicine in the last decade [6–8]. Many existing genomic prediction methods rely on either linear mixed models (LMMs) or sparse regression models. Common examples include the genomic best linear unbiased predictor (GBLUP) [9–12] and Bayesian Alphabet methods [13–16]. LMMs and sparse regression models are based on almost diametrically opposed assumptions. Precisely,

LMMs assume that all genetic variants have nonzero effect sizes and their effect sizes follow a normal distribution, whereas the sparse regression models assume that a relatively small proportion of variants affects the phenotype. Several methods considering a hybrid of the two assumptions have also been developed, and these methods combined the advantages of both LMMs and sparse regression models [17,18]. In addition, a machine-learning-based method named KAML was developed by taking full advantage of the efficient computing of LMMs and the accurate prediction of Bayesian methods [19]. However, most of these standard methods are based on statistical considerations and often ignore the prior knowledge of biological information, such as functional annotation, pathways, eQTL, and the known causal loci. Ignoring biological information is likely suboptimal, as studies have shown that incorporating the gene annotation coming from public databases or previous GWAS results can improve genomic prediction accuracy [20–24].

Our main application of interest is in genomic selection of buffalos (*Bubalus bubalis*), which is a key species for smallholder producers in developing countries (e.g., India, Pakistan, and China) [25–30] and an important milk resource for specialized markets. Parallel with the statistical methodological development, new SNP arrays for animal programs have also been developed. For example, recently, the 90K Axiom Buffalo SNP Array was designed and commercialized. Like other livestock high-density SNP chips, the 90K array for buffalo is expected to improve and speed up various genomic analyses, which include exploring genetic diversity, analyzing complex traits and diseases, and aiding genetic selection [25–30]. River-type buffalo have also been genetically selected for milk production and fertility traits in some countries by traditional methods. It has been shown that milk yield, milk components, and milk somatic cell counts have enough genetic variation for selection purposes [31–37]. Here, we ask an important question: Can we use genomic prediction to speed up the genetic gains for the buffalo population? While there are different buffalo production systems around the world, the production records are particularly limited, especially when compared to those of cow. In addition, buffalo genome research is lagging behind cow genome research, and existing approaches have to align buffalo SNPs to the bovine genome for further study [28]. Therefore, we ask the second question: Can we improve the genomic prediction performance for buffalo milk traits with the limited sample size by incorporating the related cattle QTLs?

Motivated by both methodological interest and the above two application questions, we propose a statistical approach to incorporate prior biological information in the widely used genomic best linear unbiased predictor (pGBLUP) to improve genomic prediction. We first simulated several scenarios to test the stability and advantages of the approach. Then, we applied the approach for the genomic prediction of buffalo milk traits by incorporating the known cattle milk trait QTL information from the animal QTL database [38].

## 2. Materials and Methods

### *2.1. Statistical Model*

The basic idea behind our approach is to fit the effect sizes of all SNPs as random effects relying on an LMM framework. We divided the SNPs into two groups based on a priori biological information, and we assumed that these two groups of SNPs have different effect sizes:

$$y = Zu + Z_1u_1 + \epsilon, \quad (1)$$

where y is an n by 1 phenotype vector, which has been standardized to have mean 0 and variance 1 to remove the intercept in the equation, Z is an n by m genotype matrix for all SNPs, $Z_1$ is an n by p genotype matrix for p SNPs that are part of Z, genotype matrix Z and $Z_1$ were standardized as suggested [39], u is an m by 1 vector of small effect sizes for all SNPs, $u_1$ is a p by 1 vector of additional effect sizes for p selected SNPs, $u \sim N\left(0, \frac{\sigma^2_{small}}{m}\right)$, $u_1 \sim N\left(0, \frac{\sigma^2_{large}}{P}\right)$, $\epsilon \sim N\left(0, \sigma^2_e\right)$. It should be noted that the SNPs in $Z_1$ have both large effects

and small effects, which can be drawn from $N\left(0, \left(\frac{\sigma^2_{small}}{m} + \frac{\sigma^2_{large}}{P}\right)\right)$ [10,17,18]. Recent studies found that some regions or genes in the genome were heritability enriched for complex traits and disease [9,40–44]. In this model, we extracted the biologically functional information (pathway annotation, specific gene expression, and GWAS loci information) from a public database (e.g., https://www.animalgenome.org/ (accessed on 10 October 2017)) to serve as priors to determine which set of SNPs belongs to $Z_1$. To determine whether the "prior information" is meaningful and promising, the fold of enrichment (**fe**) was used as in the stratified LD score regression or MQS [42,43]:

$$\mathbf{fe} = \left(\sigma^2_{large} / \left(\sigma^2_{large} + \sigma^2_{small}\right)\right) / (p/m) + 1. \quad (2)$$

When **fe** > 1, the genome region based on "prior information" is suggested to be heritability enriched for the complex traits and disease, and SNPs with biological information will tend to have a large effect size, while SNPs without biological information will tend to have a small effect size [42,43]. Different from the Bayesian Alphabet methods [13–16,20], which use a time-consuming MCMC algorithm to determine the SNP effect distribution, our approach directly designs the SNP effect distribution based on the prior biological information. We name this approach as the incorporating prior biological information in genomic best linear unbiased predictor (pGBLUP). Equation (1) can be written as:

$$y = g_s + g_l + \epsilon, \quad (3)$$

where $g_s = Zu$, $g_l = Z_1u_1$, $g_s \sim MVN(0, A\sigma^2_{small})$ , $g_l \sim MVN\left(0, A_1\sigma^2_{large}\right)$, $MVN$ denotes the multivariate normal distribution, and $A$ and $A_1$ are the realized genetic relationship matrix (GRM) [39]. $\sigma^2_{small}$, $\sigma^2_{large}$, and $\sigma^2_e$ are estimated firstly. Then, $g_s$ and $g_l$ can be estimated as:

$$\begin{aligned} \hat{g}_s &= A\hat{\sigma}_{small}V^{-1}y, \\ \hat{g}_l &= A_1\hat{\sigma}_{large}V^{-1}y, \end{aligned} \quad (4)$$

where $V = A\hat{\sigma}^2_s + A_1\hat{\sigma}^2_1 + I\hat{\sigma}^2_e$, $I$ is an n by n identity matrix, and the SNP effect sizes can be estimated as

$$\begin{aligned} \hat{u} &= Z^T\left(ZZ^T\right)^{-1}\hat{g}_s, \\ \hat{u}_1 &= Z_1^T\left(Z_1Z_1^T\right)^{-1}\hat{g}_l. \end{aligned} \quad (5)$$

In the training dataset, about 80% of all individuals in the simulations, the SNP effects $\hat{u_1}$ and $\hat{u}$ are estimated jointly. In the test dataset, about 20% of all individuals, the phenotypes can be predicted as:

$$\hat{y} = Z^{test}\hat{u} + Z_1^{test}\hat{u}_1 \quad (6)$$

where $Z^{test}$ is the standardized genotype matrix for all SNPs, $Z_1^{test}$ is the standardized genotype matrix for the p SNPs in the test population, and $\hat{y}$ is the predicted values, known as the genomic estimated breeding value (GEBV).

### *2.2. Animal Resources and Genomic Information*

German Holstein genomic prediction population [22,45]: The genotype data consisted of 5024 samples and 42,551 SNPs after removing SNPs that had a Hardy–Weinberg equilibrium (HWE) $p$-value < $10^{-4}$, genotype call rate < 95%, or minor allele frequency (MAF) < 0.01. All SNP positions were re-coded by the provider for confidentiality, and the genotypes of the population were used to simulate the phenotype in this study.

Water buffalo data [46]: The genotype data consisted of 412 Italian Mediterranean buffaloes, which were genotyped by the 90K Axiom Buffalo SNP Array. Then, 60,387 SNPs were retained after removing SNPs that had an HWE $p$-value < $10^{-5}$, genotype call rate < 97%, or MAF < 0.05. Six buffalo milk traits (peak milk yield (PM), total milk yield

(MY), fat yield (FY), fat percentage (FP), protein yield (PY), and protein percentage (PP)) were recorded and adjusted to 270 days in milk, as suggested by [47]. The estimated breeding value (EBV) for the six traits was estimated with a univariate animal model using ASReml 3.0 [48]. The deregressed EBV (DEBV) of the six milk production traits was calculated according to [49]. The details of the data processing were described by Liu et al. [46]. Both the EBV and DEBV were used as the phenotype in this study.

Cattle milk QTLs: The QTLs of 112 cattle-milk-related traits were downloaded from the animal QTL database (https://www.animalgenome.org/cgi-bin/QTLdb/index accessed on 10 October 2017) during October 2017; those traits included milk yield, milk fat, milk protein, and some other milk component traits. Based on the Bos taurus UMD3.1 genomic assembly, the genes within 50 kb of the QTL regions were selected as genes associated with the trait. We only focused on the QTL regions with a length smaller than 40kb, which had the largest proportion of QTLs. After initial filtering, 2435 genes were selected to be associated with the milk traits, including some genes associated with several milk traits. Then, we only kept 396 genes (see Table S1) associated with at least four traits as the prior biological information for the following real data application study for the genomic prediction of buffalo milk traits.

### 2.3. Simulations

We used the real genotypes of the German Holstein genomic prediction population to simulate the phenotypes with the following steps:

(1) Set the causal segments: The genotype matrix was standardized, and the 42,551 SNPs were divided into 1000 approximately equally sized segments, with 42 or 43 SNPs in each segment; *s* (10/25/50/100/500) segments were randomly selected as causal segments in our simulation settings, and the 10 SNPs in the center of each segment were then selected as causal SNPs; thus, the total number of causal SNPs (k) was 100/250/500/1000/5000, while the total number of SNPs in the causal segments was $p \approx s*42.5$.

(2) Simulate the SNP effects and phenotype: Firstly, all SNPs were simulated with the small effects following a normal distribution $N(0,\ 0.25/42,551)$; the k causal SNPs were simulated with additional effects following a normal distribution $N(0,\ 0.25/k)$. Then, the residual errors were sampled from a normal distribution $N(0,\ 0.5)$, so that the total heritability of the simulated trait was 0.5. Based on Equation (1), for each individual, the phenotype was obtained as the summation of small effects, large effects, and the residual error.

(3) Five-fold cross-validation: The 5024 individuals were divided into five groups, with 1004 or 1005 individuals in each group. Each time, one group of individuals was set as the test dataset, while the rest of the groups of individuals were set as the training dataset (i.e., five-fold cross-validation). We applied the pGBLUP approach in two ways to predict the performance in the test dataset: only the SNPs in the causal segments were set in $Z_1$; SNPs in both the causal segments and non-causal segments were selected in $Z_1$. We also applied the traditional GBLUP method [3] and the BayesR method [16] to compare the performance. The GBLUP method assumes the effect size for every variant is sampled from the same normal distribution; the BayesR method uses an MCMC algorithm to estimate variant effects, which are modelled as a mixture distribution of four normal distributions, including a null distribution, $N\left(0,\ 0.0\sigma_g^2\right)$, and three others: $N\left(0,\ 0.0001\sigma_g^2\right)$, $N\left(0,\ 0.001\sigma_g^2\right)$, and $N\left(0,\ 0.01\sigma_g^2\right)$, where $\sigma_g^2$ is the additive genetic variance for the trait.

### 2.4. Genomic Prediction of Buffalo Milk Traits

As the draft genomic sequence of the buffalo is currently not assigned to chromosomes, the chromosome and position for all SNPs in the 90K Axiom Buffalo SNP Array were based on the bovine UMD 3.1 genome sequence [28]. This also facilitated the use of the bovine

gene annotation information; 1279 SNPs were selected within 10 kb of the 396 cattle-milk-trait-associated genes and set in $Z_1$ of the pGBLUP model. The **fe** of "prior information" for each trait was estimated using all individuals. Then, we applied three methods, the GBLUP, BayesR, and pGBLUP, to perform the genomic prediction of the six buffalo milk traits with five-fold cross-validation: the 412 individuals were divided into five groups, with 82 or 83 individuals in each group; each time, one group of individuals was set as the test dataset, while the rest of the groups of individuals were set as the training dataset.

### *2.5. Computation*

For the GBLUP and pGBLUP, we used the GCTA [39] to perform REML, **fe** estimation, and the BLUP (http://cnsgenomics.com/software/gcta/#Download (accessed on 10 October 2017)). For BayesR, we used BayesR with the default parameters (https://github.com/syntheke/BayesR (accessed on 10 October 2017)). For the data cleaning and processing, PLINK (https://www.cog-genomics.org/plink2 (accessed in October 2017)) [50] and R (https://www.r-project.org/ (accessed on 10 October 2017)) [51] were applied.

## 3. Results

### *3.1. Predictive Accuracy in Simulations*

In our simulation studies, the causal regions were simulated and known. Firstly, we assessed the performance of different methods when only the SNPs in the causal segments were set in $Z_1$. Based on the correlation (0.45~0.56) between the predicted values and the simulated values in the test dataset, BayesR and the pGBLUP greatly outperformed the GBLUP in the sparse simulation settings (i.e., only 10 causal segments) (Figure 1A,B). If the simulated SNP effects tended to be polygenic (i.e., 500 causal segments), the three methods had a similar performance, and the GBLUP even slightly outperformed BayesR (Figure 1A,B). It should be noted that the pGBLUP had the best predictive performance in all simulation settings, especially in sparse settings in which only several causal segments greatly affected the phenotype and the causal segments were accounted for in $Z_1$. The performance of the pGBLUP depended on the **fe** of the SNP set in $Z_1$; when the **fe** decreased (Figure 1C), the predictive accuracy gain compared with the GBLUP would be reduced from about 20% to 1% (Figure 1B). The **fe** estimations were centered on the truth at the median simulated **fe**, while they tended to be underestimated at the large simulated **fe** (Figure 1C).

Then, we assessed the performance of the pGBLUP when SNPs in both the causal segments and non-causal segments were set in $Z_1$ to examine the predictability of our method. We selected 430 SNPs from different numbers of causal segments and non-causal segments and set them in $Z_1$ when simulating 10 causal segments (Figure 2). When all SNPs in $Z_1$ were from the non-causal segments, the pGBLUP and GBLUP had similar performance. Upon increasing the number of SNPs in $Z_1$ from causal segments, the predictability of the pGBLUP became better and better (Figure 2A) and the **fe** of the SNPs set in $Z_1$ also became larger and larger (Figure 2B). The pGBLUP would have the best performance when all and only the causal segments were accounted for in $Z_1$.

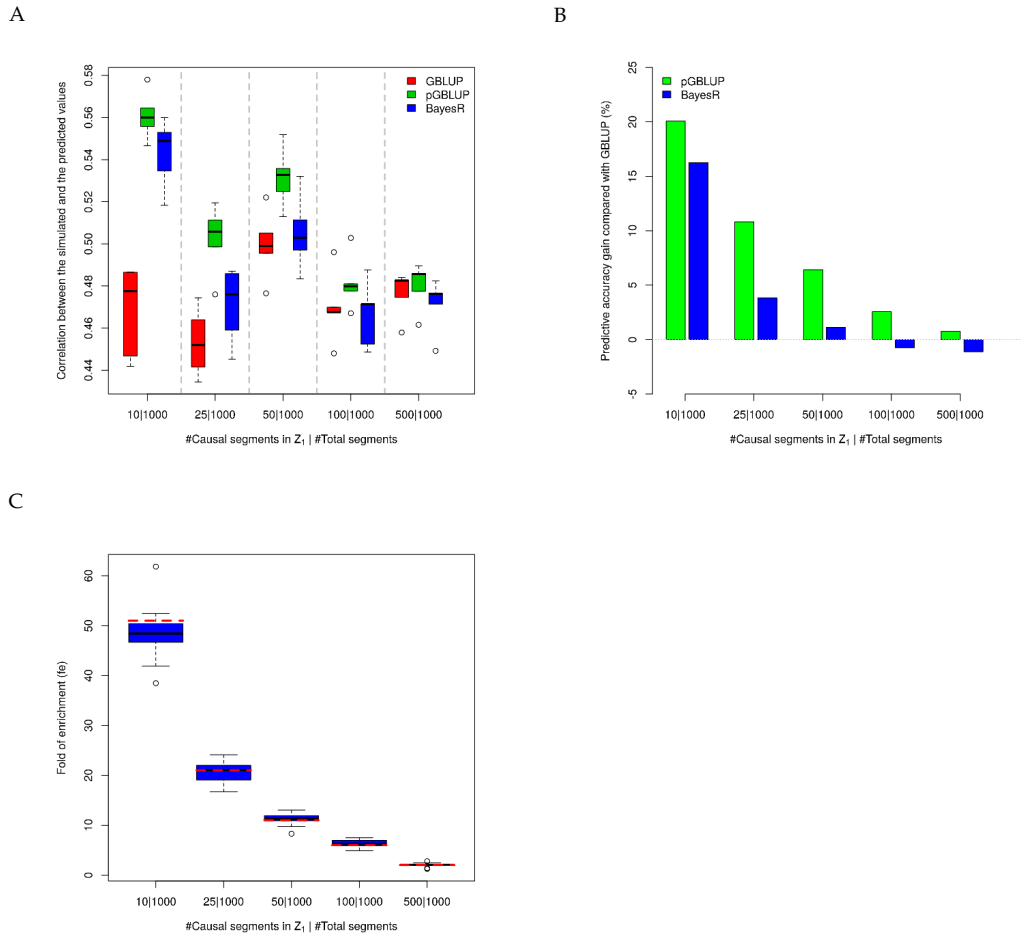


**Figure 1.** Simulation results when only the SNPs in the causal segments are set in $Z_1$. (**A**) The correlation between the predicted values and the simulated values in the test dataset using three methods in different simulation settings. (**B**) The percentage of predictive accuracy gain for the pGBLUP and BayesR compared with the GBLUP. (**C**) Fold of enrichment (**fe**) estimations using all individuals with 20 replicates; the red dashed lines represent the true values (from left to right: 51, 21, 11, 6, and 2). The black solid lines in (**A**,**C**) represent the median values of the estimations.

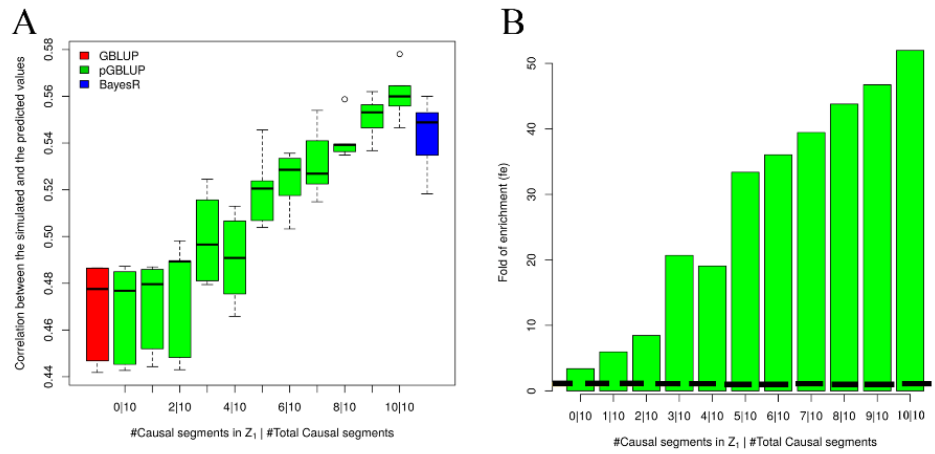


**Figure 2.** Simulation results when t = 0\1\2\3\4\5\6\7\8\9\10 causal segments and (10-t) non-causal segments are set in $Z_1$ at 10 simulated causal segments. (**A**) The correlation between the predicted values and the simulated values in the test dataset. (**B**) Average fold of enrichment (**fe**) estimations using the training dataset in the five-fold cross-validation; the black dashed line is **fe** = 1.

### *3.2. Genomic Prediction of Buffalo Milk Traits*

In our real data application studies, according to the prior biological information of 396 cattle-milk-trait-associated genes, 1279 SNPs within 10 kb of those genes were set in $Z_1$ of the pGBLUP model. The **fe** estimations of the selected 1279 SNPs in $Z_1$ are shown in Figure 3A. When the DEBV was regarded as the phenotype, there was no obvious enrichment for the selected SNPs, and the 1279 SNPs had **fe** close to 1 as other SNPs. When the EBV was regarded as the phenotype, the 1279 SNPs had a small **fe** for traits FY270, MY270, PY270, and PM.

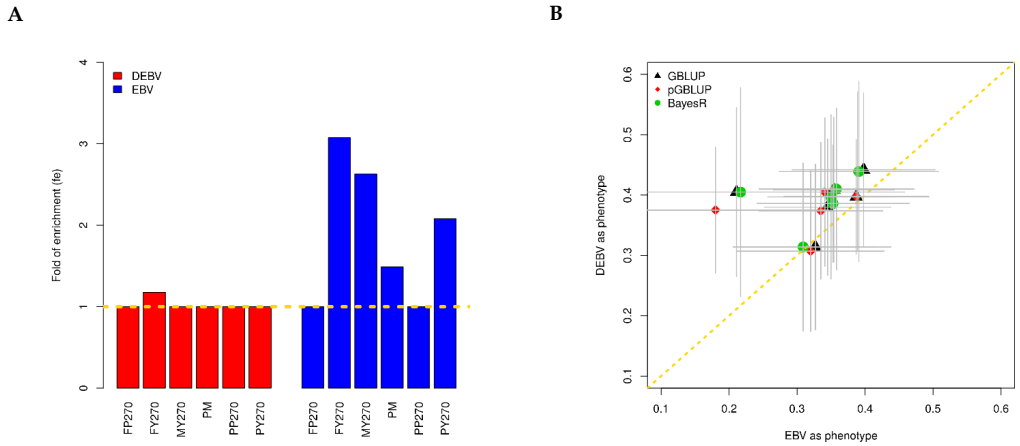


**Figure 3.** Genomic prediction results for buffalo milk traits. (**A**) Fold of enrichment (**fe**) of the selected 1279 SNPs in $Z_1$ using the DEBV and EBV as the phenotypes; the gold dashed line is **fe** = 1. (**B**) Genomic prediction performance using the EBV and DEBV as the phenotypes; the points represent the mean correlations between the GEBV and EBV (or DEBV), and the lines represent the standard errors. PM: peak milk yield; MY270: 270-day total milk yield; FY270: 270-day fat yield; FP270 = 270-day fat percentage; PY270: 270-day protein yield; PP270: 270-day protein percentage.

Three methods were applied for genomic prediction of buffalo milk traits using the GEBV and EBV as the phenotypes (Table 1). The heritability estimations of the buffalo milk traits ranged from 0.702 to 0.793 using the DEBV as the phenotype and from 0.599 to 0.741 using the EBV as the phenotype. The correlations between the GEBV and DEBV ranged from 0.304 to 0.442, while the correlations between the GEBV and EBV ranged from 0.180 to 0.398. The three methods had similar performance for genomic prediction with large standard errors, and using the DEBV as the phenotype had better genomic prediction performance than using the EBV as the phenotype on average (Figure 3B and Table 1).

**Table 1.** Genomic prediction results of buffalo milk traits using the GEBV and EBV as the phenotypes. The mean and standard error of the correlation between the predicted and the true values in the five-fold cross-validation are reported. PM: peak milk yield; MY270: 270-day total milk yield; FY270: 270-day fat yield; FP270 = 270-day fat percentage; PY270: 270-day protein yield; PP270: 270-day protein percentage.

| | Trait | $h^2$ | GBLUP | pGBLUP | BayesR |
|---|---|---|---|---|---|
| DEBV | FP270 | 0.713 ± 0.112 | 0.314 ± 0.137 | 0.307 ± 0.133 | 0.314 ± 0.139 |
| | FY270 | 0.703 ± 0.119 | 0.38 ± 0.113 | 0.374 ± 0.113 | 0.397 ± 0.136 |
| | MY270 | 0.753 ± 0.112 | 0.409 ± 0.12 | 0.405 ± 0.123 | 0.41 ± 0.134 |
| | PM | 0.702 ± 0.115 | 0.405 ± 0.14 | 0.375 ± 0.104 | 0.405 ± 0.173 |
| | PP270 | 0.75 ± 0.112 | 0.397 ± 0.095 | 0.398 ± 0.092 | 0.386 ± 0.097 |
| | PY270 | 0.793 ± 0.108 | 0.442 ± 0.127 | 0.439 ± 0.132 | 0.439 ± 0.149 |
| EBV | FP270 | 0.741 ± 0.114 | 0.327 ± 0.111 | 0.32 ± 0.108 | 0.309 ± 0.103 |
| | FY270 | 0.631 ± 0.124 | 0.345 ± 0.093 | 0.335 ± 0.091 | 0.35 ± 0.093 |
| | MY270 | 0.658 ± 0.122 | 0.354 ± 0.089 | 0.341 ± 0.077 | 0.358 ± 0.114 |
| | PM | 0.599 ± 0.123 | 0.211 ± 0.22 | 0.18 ± 0.167 | 0.217 ± 0.241 |
| | PP270 | 0.726 ± 0.115 | 0.387 ± 0.107 | 0.387 ± 0.106 | 0.353 ± 0.112 |
| | PY270 | 0.738 ± 0.116 | 0.398 ± 0.105 | 0.389 ± 0.101 | 0.391 ± 0.117 |

## 4. Discussion

We proposed a new genomic prediction approach called the pGBLUP, which incorporates prior biological information in the LMM. Several methods incorporating prior biological information in the LMM were also developed recently using different strategies from ours: The BLUP | GA [21,22,52,53] and single-step GBLUP accounting for causative quantitative trait nucleotides (QTNs) [24] model one weighted trait-specific GRM based on the prior biological information; CVAT [23,54,55] models two genetic variances in the LMM, while the SNPs in the two genetic components should be disjointly divided by the prior biological information. Both the BLUP | GA and CVAT selected and tested prior biological information using some iteration or permutation procedures, while we followed the main idea from [42,43] and directly used the heritability enrichment (**fe**) to measure the importance of the prior biological information. As our simulation results show in Figures 1 and 2, the predictability of pGBLUP could be improved as the **fe** of prior biological information increased. Under the Bayesian framework, the extension of BayesR, BayesRC, was introduced, which incorporates prior biological information in the analysis by defining classes of variants likely to be enriched for SNPs with prior biological information, which showed competitive performance in the QTL mapping and genomic predictions [20].

In our simulations, the real genotypes of the cattle were divided into 1000 segments without considering the linkage disequilibrium (LD), which is not ideal. However, this should not affect our simulation purpose of illustrating the relationship between the predictive performance and **fe**; the pGBLUP and BayesR performed better with larger **fe**, which indicated less genes had large effects on the traits, and the GBLUP performed better with smaller **fe**, indicating more genes had small effects on the traits (Figure 1). While ignoring the LD in the simulations may affect the **fe** estimation, when there were 10 causal segments, the **fe** of all SNPs in the 10 causal segments was underestimated (Figure 1C). The missing **fe** were shared by the nearby non-causal segments, which could be observed for the **fe** overestimation when all SNPs from the non-causal segments had **fe** $> 1$ (Figure 2B). In summary, the **fe** estimation was a good measurement for the importance and quality of the prior biological information. If the prior biological information was appropriate, the pGBLUP would outperform BayesR and the GBLUP; otherwise, it performed slightly worse than BayesR and equal to or better than the GBLUP (Figure 2A).

We applied the pGBLUP to our published buffalo data [46] with another two popular genomic prediction methods: the GBLUP and BayesR. Due to the delayed buffalo genome research, prior biological information is very rare. For buffalo traits, some cattle-related QTLs were also identified in buffalo [25,28,56–59], while the sample size for the buffalo population was too small to detect more causal variants. In this study, we incorporated the known cattle milk trait QTLs [38] for genomic prediction of buffalo milk traits in the pGBLUP. The prior biological information borrowed from cattle showed median enrichment for FY270, MY270, PM, and PY270 using the EBV as the phenotype. If the DEBV was used as the phenotype, the prior biological information only showed small enrichment for FY270, which had the largest enrichment when using the EBV as the phenotype (Figure 3). The **fe** estimations suggested that the prior information has the potential to improve the genomic predictability for FY270, MY270, PM, and PY270 if the EBV was used as the phenotype and for FY270 if the DEBV was used as the phenotype. The predictabilities of the three methods did not vary much (Table 1); BayesR and the pGBLUP did not show an advantage in genomic prediction, indicating the buffalo milk traits were less artificially selected for genes with large effects compared with cattle [13,16,52]. The other reasons for the small difference may be due to the limited sample size of the buffalo population in this study, which was indicated by the large standard errors of the correlations. In addition, we used a relatively loose threshold for the HWE test ($p$-value $< 10^{-5}$) to remove the variants due to the genotyping errors, which was also likely to remove the causal variants under selection, thus affecting the performance of the genomic prediction. The heritability estimations using the DEBV and EBV as the phenotypes were larger than those using the original records directly [37,60,61], because some environmental effects were already removed for the DEBV and EBV. As previous research, we

also noticed that the DEBV as the response variable yielded more reliable genomic predictions than the traditional EBV [49,62]. When the DEBV was used as the phenotype, the heritability estimations could reach 0.793 for PY270; the maximum achievable correlation between the predicted and observed traits was $\sqrt{0.793} = 0.891$, but the mean correlations using the GBLUP, pGBLUP, and BayesR were 0.442, 0.439, and 0.439 (Table 1), so there was a large gap between the maximum achievable correlation and the real correlation. Based on the simulation results that the simulated heritability was 0.5 and the maximum achievable correlation was $\sqrt{0.5} = 0.707$, the realistic correlation could reach 0.58 (Figure 1A); we believe that the genomic prediction of buffalo traits can be further improved with a larger sample size, higher-density SNP chips, and more precise prior biological information.

## 5. Conclusions

We proposed a genomic prediction approach, the pGBLUP, which has the potential to improve the genomic prediction performance by incorporating the proper prior biological information. The pGBLUP uses heritability enrichment to quickly check the importance of the prior biological information. We also applied the pGBLUP to incorporate the milk-related QTL information from cows for genomic prediction of buffalo milk traits. We found that some cattle-milk-related QTLs also played an important role in buffalo milk production traits. We believe that genomic prediction of buffalo traits can be further improved with a larger sample size, higher-density SNP chips, and more precise prior biological information.

**Supplementary Materials:** The following Supporting Information can be downloaded at: https://www.mdpi.com/article/10.3390/genes13081430/s1, Table S1: The gene list in the QTL of cattle milk traits.

**Author Contributions:** X.H., S.Z. and L.Y. conceived of and designed the experiments; S.Z. and L.Y. supervised the study; L.Y. and G.P. provided the regents; A.S., B.G., G.C. and A.L. collected the buffalo samples and records; X.H., A.L. and J.L. analyzed the data; J.L., C.Z. and Z.W. performed the experiments; X.H. and A.L. wrote the manuscript, Z.W. critically reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This project was supported by the earmarked fund for CARS 36 and the International Cooperation Key Project of China (No. 2011 DFA32250).

**Institutional Review Board Statement:** No experimental animal studies were conducted for the work detailed in this manuscript. References have been provided where animal data were used.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The production data of buffalo used in this study were provided by The Italian Buffalo Breeders Association (ANASB), which is responsible for the official herd book of the buffalo population in Italy. The buffalo data are available from the authors upon reasonable request. The cattle data were provided by Zhe Zhang from the South China Agricultural University (zhezhang@scau.edu.cn), which are available from Zhe Zhang upon reasonable request.

**Acknowledgments:** We thank Zhe Zhang from the South China Agricultural University for providing us with the cattle data and Xiang Zhou from University of Michigan for the helpful discussion.

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

## Abbreviations

| | |
|---|---|
| QTL | quantitative trait locus |
| QTN | quantitative trait nucleotide |
| SNP | single-nucleotide polymorphism |
| GBLUP | genomic best linear unbiased predictor |
| pGBLUP | incorporating prior biological information in genomic best linear unbiased predictor |
| EBV | estimated breeding value |
| GEBV | genomic estimated breeding value |
| DEBV | deregressed estimated breeding value |
| LMM | linear mixed models |
| fe | fold of enrichment |

## References


1. Hickey, J.M.; Chiurugwi, T.; Mackay, I.; Powell, W. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* **2017**, *49*, 1297. [CrossRef]
2. Goddard, M.E.; Hayes, B.J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* **2009**, *10*, 381–391. [CrossRef] [PubMed]
3. Meuwissen, T.H.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829. [CrossRef] [PubMed]
4. Liu, X.; Huang, M.; Fan, B.; Buckler, E.S.; Zhang, Z. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLoS Genet.* **2016**, *12*, e1005767. [CrossRef] [PubMed]
5. Georges, M.; Charlier, C.; Hayes, B. Harnessing genomic information for livestock improvement. *Nat. Rev. Genet.* **2019**, *20*, 135–156. [CrossRef]
6. Lee, S.H.; Weerasinghe, W.S.P.; Wray, N.R.; Goddard, M.E.; Van Der Werf, J.H. Using information of relatives in genomic prediction to apply effective stratified medicine. *Sci. Rep.* **2017**, *7*, 42091. [CrossRef] [PubMed]
7. Chatterjee, N.; Shi, J.; García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **2016**, *17*, 392–406. [CrossRef]
8. de los Campos, G.; Vazquez, A.I.; Fernando, R.; Klimentidis, Y.C.; Sorensen, D. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. *PLoS Genet.* **2013**, *9*, e1003608. [CrossRef] [PubMed]
9. Jensen, J.; Su, G.; Madsen, P. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. *BMC Genet.* **2012**, *13*, 44. [CrossRef]
10. Coram, M.A.; Fang, H.; Candille, S.I.; Assimes, T.L.; Tang, H. Leveraging Multi-ethnic Evidence for Risk Assessment of Quantitative Traits in Minority Populations. *Am. J. Hum. Genet.* **2017**, *101*, 218–226. [CrossRef] [PubMed]
11. Speed, D.; Balding, D.J. MultiBLUP: Improved SNP-based prediction for complex traits. *Genome Res.* **2014**, *24*, 1550–1557. [CrossRef]
12. Da, Y.; Wang, C.; Wang, S.; Hu, G. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS ONE* **2014**, *9*, e87666. [CrossRef]
13. Kemper, K.E.; Reich, C.M.; Bowman, P.J.; Vander Jagt, C.J.; Chamberlain, A.J.; Mason, B.A.; Hayes, B.J.; Goddard, M.E. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genet. Sel. Evol.* **2015**, *47*, 29. [CrossRef]
14. Habier, D.; Fernando, R.L.; Kizilkaya, K.; Garrick, D.J. Extension of the bayesian alphabet for genomic selection. *BMC Bioinform.* **2011**, *12*, 186. [CrossRef]
15. Moser, G.; Lee, S.H.; Hayes, B.J.; Goddard, M.E.; Wray, N.R.; Visscher, P.M. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet.* **2015**, *11*, e1004969. [CrossRef]
16. Erbe, M.; Hayes, B.; Matukumalli, L.K.; Goswami, S.; Bowman, P.J.; Reich, C.M. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* **2012**, *95*, 4114–4129. [CrossRef]
17. Zhou, X.; Carbonetto, P.; Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **2013**, *9*, e1003264. [CrossRef]
18. Zeng, P.; Zhou, X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat. Commun.* **2017**, *8*, 456. [CrossRef]
19. Yin, L.; Zhang, H.; Zhou, X.; Yuan, X.; Zhao, S.; Li, X.; Liu, X. KAML: Improving genomic prediction accuracy of complex traits using machine learning determined parameters. *Genome Biol.* **2020**, *21*, 146. [CrossRef]
20. 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 Genom.* **2016**, *17*, 144. [CrossRef]
21. Gao, N.; Martini, J.W.R.; Zhang, Z.; Yuan, X.; Zhang, H.; Simianer, H.; Li, J. Incorporating Gene Annotation into Genomic Prediction of Complex Phenotypes. *Genetics* **2017**, *207*, 489–501. [CrossRef]
22. Zhang, Z.; Erbe, M.; He, J.; Ober, U.; Gao, N.; Zhang, H.; Simianer, H.; Li, J. Accuracy of Whole-Genome Prediction Using a Genetic Architecture-Enhanced Variance-Covariance Matrix. *G3 Genes Genomes Genet.* **2015**, *5*, 615–627. [CrossRef] [PubMed]
23. Fang, L.; Sahana, G.; Ma, P.; Su, G.; Yu, Y.; Zhang, S.; Lund, M.S.; Sørensen, P. Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet. Sel. Evol.* **2017**, *49*, 44. [CrossRef]
24. Fragomeni, B.O.; Lourenco, D.A.L.; Masuda, Y.; Legarra, A.; Misztal, I. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. *Genet. Sel. Evol.* **2017**, *49*, 59. [CrossRef] [PubMed]
25. El-Halawany, N.; Abdel-Shafy, H.; Shawky, A.-E.-M.A.; Abdel-Latif, M.A.; Al-Tohamy, A.F.M.; Abd El-Moneim, O.M. Genome-wide association study for milk production in Egyptian buffalo. *Livest. Sci.* **2017**, *198* (Suppl. C), 10–16. [CrossRef]
26. Michelizzi, V.N.; Wu, X.; Dodson, M.V.; Michal, J.J.; Zambrano-Varon, J.; McLean, D.J.; Jiang, Z. A Global View of 54,001 Single Nucleotide Polymorphisms (SNPs) on the Illumina BovineSNP50 BeadChip and Their Transferability to Water Buffalo. *Int. J. Biol. Sci.* **2011**, *7*, 18–27. [CrossRef]

27. Iamartino, D.; Williams, J.L.; Sonstegard, T.; Reecy, J.; Tassell Cv Nicolazzi, E.L.; Biffani, S.; Biscarini, F.; Schroeder, S.; de Oliveira, D.A. The buffalo genome and the application of genomics in animal management and improvement. *Buffalo Bull.* **2013**, *32*, 151–158.
28. Iamartino, D.; Nicolazzi, E.L.; Van Tassell, C.P.; Reecy, J.M.; Fritz-Waters, E.R.; Koltes, J.E.; Biffani, S.; Sonstegard, T.S.; Schroeder, S.G.; Ajmone-Marsan, P. Design and validation of a 90K SNP genotyping assay for the water buffalo (Bubalus bubalis). *PLoS ONE* **2017**, *12*, e0185220. [CrossRef] [PubMed]
29. De Camargo, G.; Aspilcueta-Borquis, R.R.; Fortes, M.; Porto-Neto, R.; Cardoso, D.F.; Santos, D.; Lehnert, S.; Reverter, A.; Moore, S.; Tonhati, H. Prospecting major genes in dairy buffaloes. *BMC Genomics* **2015**, *16*, 872. [CrossRef] [PubMed]
30. Aspilcueta-Borquis, R.; Neto, F.A.; Santos, D.; Hurtado-Lugo, N.; Silva, J.; Tonhati, H. Multiple-trait genomic evaluation for milk yield and milk quality traits using genomic and phenotypic data in buffalo in Brazil. *Gen. Mol. Res.* **2015**, *14*, 18009–18017. [CrossRef] [PubMed]
31. Borquis, R.R.A.; de Araujo Neto, F.R.; Baldi, F.; Hurtado-Lugo, N.; de Camargo, G.M.; Muñoz-Berrocal, M.; Tonhati, H. Multiple-trait random regression models for the estimation of genetic parameters for milk, fat, and protein yield in buffaloes. *J. Dairy Sci.* **2013**, *96*, 5923–5932. [CrossRef]
32. Hossein-Zadeh, N.G.; Nazari, M.A.; Shadparvar, A.A. Genetic perspective of milk yield persistency in the first three lactations of Iranian buffaloes (Bubalus bubalis). *J. Dairy Res.* **2017**, *84*, 434–439. [CrossRef] [PubMed]
33. Patil, H.R. *Genetic Evaluation of Fertility and Production Efficiency Traits in Murrah Buffalo*; LUVAS: Hisar, India, 2016.
34. Agudelo-Gómez, D.; Pelicioni Savegnago, R.; Buzanskas, M.; Ferraudo, A.; Prado Munari, D.; Cerón-Muñoz, M. Genetic principal components for reproductive and productive traits in dual-purpose buffaloes in Colombia. *J. Anim. Sci.* **2015**, *93*, 3801–3809. [CrossRef] [PubMed]
35. Dash, S.; Chakravarty, A.; Singh, A.; Shivahre, P.R.; Upadhyay, A.; Sah, V.; Singh, K.M. Assessment of expected breeding values for fertility traits of Murrah buffaloes under subtropical climate. *Vet. World* **2015**, *8*, 320. [CrossRef]
36. Gupta, J.P.; Sachdeva, G.K.; Gandhi, R.; Chakaravarty, A. Developing multiple-trait prediction models using growth and production traits in Murrah buffalo. *Buffalo Bull.* **2015**, *34*, 347–355.
37. Aspilcueta-Borquis, R.; Neto, F.A.; Baldi, F.; Bignardi, A.; Albuquerque, L.G.; Tonhati, H. Genetic parameters for buffalo milk yield and milk quality traits using Bayesian inference. *J. Dairy Sci.* **2010**, *93*, 2195–2201. [CrossRef] [PubMed]
38. Hu, Z.-L.; Park, C.A.; Reecy, J.M. Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Res.* **2016**, *44*, D827–D833. [CrossRef]
39. Yang, J.; Lee, S.H.; Goddard, M.E.; Visscher, P.M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **2011**, *88*, 76–82. [CrossRef]
40. Edwards, S.M.; Thomsen, B.; Madsen, P.; Sorensen, P. Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle. *Genet. Sel. Evol.* **2015**, *47*, 60. [CrossRef]
41. Gusev, A.; Lee, S.H.; Trynka, G.; Finucane, H.; Vilhjálmsson, B.J.; Xu, H. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **2014**, *95*, 535–552. [CrossRef]
42. Finucane, H.K.; Bulik-Sullivan, B.; Gusev, A.; Trynka, G.; Reshef, Y.; Loh, P.R.; Anttila, V. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **2015**, *47*, 1228–1235. [CrossRef] [PubMed]
43. Zhou, X. A Unified Framework for Variance Component Estimation with Summary Statistics in Genome-wide Association Studies. *Ann Appl Stat.* **2017**, *11*, 2027–2051. [CrossRef] [PubMed]
44. Fang, L.; Sahana, G.; Su, G.; Yu, Y.; Zhang, S.; Lund, M.S.; Sørensen, P. Integrating Sequence-based GWAS and RNA-Seq Provides Novel Insights into the Genetic Basis of Mastitis and Milk Production in Dairy Cattle. *Sci. Rep.* **2017**, *7*, 45560. [CrossRef] [PubMed]
45. Matukumalli, L.K.; Lawley, C.T.; Schnabel, R.D.; Taylor, J.F.; Allan, M.F.; Heaton, M.P. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE* **2009**, *4*, e5350. [CrossRef]
46. Liu, J.J.; Liang, A.X.; Campanile, G.; Plastow, G.; Zhang, C.; Wang, Z.; Salzano, A.; Gasparrini, B.; Cassandro, M.; Yang, L.G. Genome-wide association studies to identify quantitative trait loci affecting milk production traits in water buffalo. *J. Dairy Sci.* **2018**, *101*, 433–444. [CrossRef] [PubMed]
47. Baldi, F.; Laureano, M.M.M.; Gordo, D.G.M.; Bignardi, A.B.; Borquis, R.R.A.; Albuquerque, L.G.; Tonhati, H. Effect of lactation length adjustment procedures on genetic parameter estimates for buffalo milk yield. *Genet. Mol. Biol.* **2011**, *34*, 62–67. [CrossRef]
48. Gilmour, A.R.; Gogel, R.B.J.; Cullis, B.R.; Thompson, R. Asreml User Guide Release 3.0; 2009. Available online: https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/2018/02/ASReml-3-User-Guide.pdf (accessed on 10 October 2017).
49. Garrick, D.J.; Taylor, J.F.; Fernando, R.L. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* **2009**, *41*, 55. [CrossRef] [PubMed]
50. 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. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef] [PubMed]
51. Ihaka, R.; Gentleman, R. R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **1996**, *5*, 299–314.
52. Zhang, Z.; Ober, U.; Erbe, M.; Zhang, H.; Gao, N.; He, J.; Li, J.; Simianer, H. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS ONE* **2014**, *9*, e93017. [CrossRef]

53. Zhang, Z.; Liu, J.; Ding, X.; Bijma, P.; de Koning, D.J.; Zhang, Q. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS ONE* **2010**, *5*, e12648. [CrossRef]
54. Rohde, P.D.; Demontis, D.; Cuyabano, B.C.D.; Børglum, A.D.; Sørensen, P. Covariance association test (CVAT) identify genetic markers associated with schizophrenia in functionally associated biological processes. *Genetics* **2016**, *203*, 1901–1913. [CrossRef]
55. Fang, L.; Sahana, G.; Ma, P.; Su, G.; Yu, Y.; Zhang, S.; Lund, M.S.; Sørensen, P. Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds. *BMC Genomics* **2017**, *18*, 604. [CrossRef] [PubMed]
56. Deng, T.; Pang, C.; Ma, X.; Duan, A.; Liang, S.; Lu, X.; Liang, X. Buffalo SREBP1: Molecular cloning, expression and association analysis with milk production traits. *Anim. Genet.* **2017**, *48*, 720–721. [CrossRef] [PubMed]
57. Dinesh, K.; Verma, A.; Gupta, I.D.; Thakur, Y.P.; Verma, N.; Arya, A. Identification of polymorphism in exons 7 and 12 of lactoferrin gene and its association with incidence of clinical mastitis in Murrah buffalo. *Trop. Anim. Health Prod.* **2015**, *47*, 643–647. [CrossRef]
58. El-Magd, M.A.; Abo-Al-Ela, H.G.; El-Nahas, A.; Saleh, A.A.; Mansour, A.A. Effects of a novel SNP of IGF2R gene on growth traits and expression rate of IGF2R and IGF2 genes in gluteus medius muscle of Egyptian buffalo. *Gene* **2014**, *540*, 133–139. [CrossRef]
59. Yuan, J.; Zhou, J.; Deng, X.; Hu, X.; Li, N. Molecular cloning and single nucleotide polymorphism detection of buffalo DGAT1 gene. *Biochem. Genet.* **2007**, *45*, 611–621. [CrossRef]
60. Rosati, A.; Van Vleck, L.D. Estimation of genetic parameters for milk, fat, protein and mozzarella cheese production for the Italian river buffalo Bubalus bubalis population. *Livest. Prod. Sci.* **2002**, *74*, 185–190. [CrossRef]
61. Malhado, C.H.M.; Malhado, A.C.M.; Ramos, A.d.A.; Carneiro, P.L.S.; Souza JCd Pala, A. Genetic parameters for milk yield, lactation length and calving intervals of Murrah buffaloes from Brazil. *Rev. Bras. De Zootec.* **2013**, *42*, 565–569. [CrossRef]
62. Ostersen, T.; Christensen, O.F.; Henryon, M.; Nielsen, B.; Su, G.; Madsen, P. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. *Genet. Sel. Evol.* **2011**, *43*, 38. [CrossRef] [PubMed]