animals

MDPI

Article

PCA-Based Multiple-Trait GWAS Analysis: A Powerful Model for Exploring Pleiotropy

Wengang Zhang [1,†], Xue Gao [1,†], Xinping Shi [1,2], Bo Zhu [1], Zezhao Wang [1], Huijiang Gao [1], Lingyang Xu [1], Lupei Zhang [1], Junya Li [1,*] and Yan Chen [1,*]

[1] Cattle Genetics and Breeding Group, Institute of Animal Science (IAS), Chinese Academy of Agricultural Sciences (CAAS), Beijing 100193, China; zhangwengang_19@sina.com (W.Z.); gaoxue76@126.com (X.G.); sxp18811727129@163.com (X.S.); zhubo525@126.com (B.Z.); wangzezhao1@163.com (Z.W.); gaohj111@sina.com (H.G.); xulingyang@caas.cn (L.X.); zhanglupei@caas.cn (L.Z.)

[2] College of Animal Science and Technology, Hebei Agricultural University, Baoding 071000, China

* Correspondence: lijunya@caas.cn (J.L.); chenyan0204@163.com (Y.C.); Tel.: +86-138-1156-8766 (J.L.); +86-134-3967-4745 (Y.C.)

† These authors contributed equally to this work.

Received: 30 August 2018; Accepted: 28 November 2018; Published: 17 December 2018

Simple Summary: In biological processes, it is common that a single gene controls two or more traits, leading to a high genetically correlation between many traits in human beings and livestock. Genome-wide association study (GWAS) is a popular method for mapping causal genes or regions related to studied traits. Taking the advantage of genetically correlation among traits, a combined analysis of two or more traits can improve the power of detection in GWAS analysis. In this study, we prove the improvement of multiple-traits GWAS through theoretical derivation, simulated dataset and real dataset, respectively. In addition, using this approach, we successfully identified a candidate gene for presoma muscle development in cattle that were not be found in the average association analysis. In summary, we conclude that multiple-trait GWAS is an effective method to explore genetic factors of traits, which have high correlations.

Abstract: Principal component analysis (PCA) is a potential approach that can be applied in multiple-trait genome-wide association studies (GWAS) to explore pleiotropy, as well as increase the power of quantitative trait loci (QTL) detection. In this study, the relationship of test single nucleotide polymorphisms (SNPs) was determined between single-trait GWAS and PCA-based GWAS. We found that the estimated pleiotropic quantitative trait nucleotides (QTNs) $\hat{\beta}^*$ were in most cases larger than the single-trait model estimations ($\hat{\beta}_1$ and $\hat{\beta}_2$). Analysis using the simulated data showed that PCA-based multiple-trait GWAS has improved statistical power for detecting QTL compared to single-trait GWAS. For the minor allele frequency (MAF), when the MAF of QTNs was greater than 0.2, the PCA-based model had a significant advantage in detecting the pleiotropic QTNs, but when its MAF was reduced from 0.2 to 0, the advantage began to disappear. In addition, as the linkage disequilibrium (LD) of the pleiotropic QTNs decreased, its detection ability declined in the co-localization effect model. Furthermore, on the real data of 1141 Simmental cattle, we applied the PCA model to the multiple-trait GWAS analysis and identified a QTL that was consistent with a candidate gene, *MCHR2*, which was associated with presoma muscle development in cattle. In summary, PCA-based multiple-trait GWAS is an efficient model for exploring pleiotropic QTNs in quantitative traits.

Keywords: genome-wide association study; principal component analysis; multiple-trait; pleiotropy; *MCHR2*

1. Introduction

Disease and quantitative traits usually follow a polygenic model [1], in which quantitative trait loci (QTL) and candidate genes can be explored using genome-wide association studies (GWAS) [2]. In general, candidate genes or causal variants can affect multiple traits simultaneously, a phenomenon known as "pleiotropy", that usually occurs when traits share common quantitative trait nucleotides (QTNs), or QTNs in traits have a high linkage disequilibrium (LD) [3]. Typical pleiotropic traits are phenotypically or genetically correlated and are unconstrained, such as disease traits, quantitative traits, and Mendelian traits. According to the National Human Genome Research Institute (NHGRI) [4], pleiotropy exists in 17% of trait-associated genes and 5% of trait-associated single nucleotide polymorphisms (SNPs). Studies on Crohn's disease and psoriasis [5], and body mass index (BMI) and melanoma [6], have highlighted numerous pleiotropic QTNs.

A plausible approach for exploring pleiotropy is the multiple-trait GWAS model in comparison with single trait GWAS, which has been shown to be an effective method to detect shared QTL [7]. Although a multivariate model with multiple traits is a powerful approach, it requires a large amount of computation time and computational memory capacity [8], because it must solve a covariance matrix of $np \times np$ in size (n, number of individuals; p, number of traits), with a time complexity of $O(n^3p^3 \cdot t)$. Some researchers [9–11] have reduced the computation time, however the multivariate model is still costly when many traits are considered together. Based on principal component analysis (PCA) and linear discriminant analysis, another powerful model utilizes dimension reduction of traits to track pleiotropy [12,13]. PCA-based multiple-trait GWAS has been shown to explain the largest amount of heritability [14], as well as to be robust and powerful in practice [15]. Compared with the multivariate model, this method takes much less time, therefore it has been widely used in pleiotropic QTL mapping [16]. However, it should be noted that one limitation of PCA-based GWAS is that it can only be applied when all traits are measured on all samples.

In livestock breeding, fine mapping of pleiotropic QTL for objective traits, such as milk yield, milk fat yield, and milk protein yield in dairy cattle [17,18], as well as the average daily gain and carcass weight in beef cattle [19], is important. Christine conducted a PCA-based multiple-trait GWAS and identified two regions (SSC5: 21.3 Mb–25.1 Mb, SSC14: 151.5 Mb–154.0 Mb) that have pleiotropic effects on boar taint components and testicular traits [20]. It helps to better understand the genetic mechanisms of complex traits, especially those related to commercial traits, and provide guidance for marker-assisted selection (MAS) in domestic animal breeding.

In this study, we considered two types of pleiotropy, namely a single causal variant model and a colocalizing effect model. Specifically, the colocalizing effect model is defined as different causal variants that affect distinguishing phenotypes with high linkage disequilibrium (LD), resulting in variants displaying signals in association with different traits. We first theoretically describe the relationship between a PCA-based multiple-trait GWAS model and single-trait model for pleiotropic QTL mapping. Next, we demonstrate a powerful PCA-based model based on three sets of simulation data under three situations (medium heritability traits, low heritability traits, and environmental correlation traits). Finally, we use real GWAS data of three meat cut traits to explore candidate genes associated with presoma development in cattle. The analytical strategies are visually outlined in Figure 1.

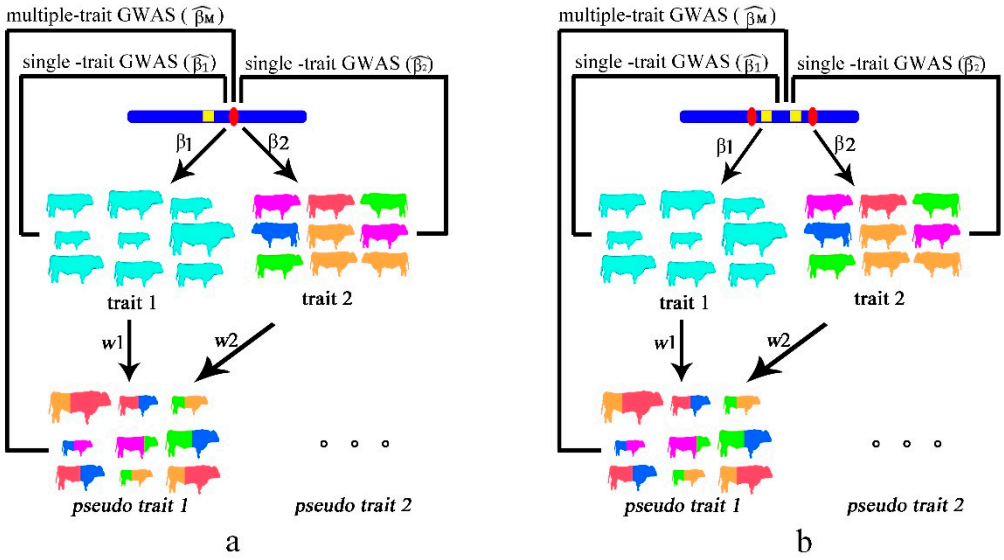

Figure 1. Layout of principal component analysis (PCA)-based multiple-trait genome-wide association studies (GWAS) versus single-trait GWAS. (**a**) Single causal variant model. Provided that a casual single nucleotide polymorphism (SNP) (red spot) has an effect on trait 1 (cattle size) and trait 2 (cattle color) with $\beta 1$ and $\beta 2$, the process of estimation of $\beta 1$ and $\beta 2$ using trait 1 and trait 2 is called single-trait GWAS. According to components decomposition, pseudo traits are formed and the process of estimation of β_M is called PCA-based multiple-trait GWAS. The yellow marker represents genotyped SNP in beadchip. (**b**) Colocalizing effect model. Two different genetic variants in high linkage disequilibrium that affect different traits. In both situations, we compared the relationships among $\beta 1$, $\beta 2$, and β_M.

2. Method

We firstly decomposed the phenotypes into several principal components scores (PCS) according to eigenvectors, and then treated PCS as pseudo traits to carry out multiple-trait GWAS. To show the improved power of PCA-based GWAS, we theoretically explored the relationship of the estimated effects between PCA-based multiple-trait GWAS and single-trait GWAS. In this study, two situations were considered as follows.

2.1. Single Causal Variant Model

In GWAS analysis, the standard approach usually uses a mixed linear model (MLM), in which polygenic effects are treated as random effects [21]. For a clearer comparison with the two association strategies (multi-traits GWAS and single-trait GWAS), we simplified the GWAS model into a general linear model (GLM) instead of a MLM (Figure 1). Here, we referred to a GLM in a QTL mapping study [22] (also called least-squares regression if only a SNP effect is considered in the model). X is the genotype matrix for a single marker, defined as 0 for the heterozygote and -1 and 1 for the two homozygotes. Two traits were observed (represented by y_1 and y_2) and included in single-marker GLM tests as follows:

$$y_1 = X\beta_1 + e_1 \quad (1)$$

$$y_2 = X\beta_2 + e_2 \quad (2)$$

where β_1 and β_2 represent the marker's effect on trait one and trait two, respectively. Therefore, β_1 and β_2 are estimated by

$$\hat{\beta_1} = \left(X^T X\right)^{-1} X^T y_1 \quad (3)$$

$$\hat{\beta_2} = \left(X^T X\right)^{-1} X^T y_2 \quad (4)$$

The phenotypes followed E(y_1) = 0 and E(y_2) = 0 after phenotype normalization. We conducted principal component analysis (PCA) between phenotypic traits in two steps. First, we constructed the covariance matrix *S*:

$$\mathrm{S} = \begin{bmatrix} \frac{(y_1-\bar{y}_1)^T(y_1-\bar{y}_1)}{n-1} & \frac{(y_1-\bar{y}_1)^T(y_2-\bar{y}_2)}{n-1} \\ \frac{(y_2-\bar{y}_2)^T(y_1-\bar{y}_1)}{n-1} & \frac{(y_2-\bar{y}_2)^T(y_2-\bar{y}_2)}{n-1} \end{bmatrix} = \frac{1}{n-1}\begin{bmatrix} y_1^T y_1 & y_1^T y_2 \\ y_2^T y_1 & y_2^T y_2 \end{bmatrix} \quad (5)$$

where *n* is the number of phenotyped individuals. Second, we created a pseudo trait weighting of the first eigenvector (μ):

$$y^* = [y_1, y_2]\mu \quad (6)$$

Therefore, the linear regression analysis and marker's effect estimation of β* can be written as

$$y^* = X\beta^* + e^* \quad (7)$$

$$\hat{\beta^*} = \left(X^T X\right)^{-1} X^T y^* \quad (8)$$

Here, we compared the pseudo trait effect (β*) with two traits effects (β_1 and β_2) to explain the increasing power using the pseudo trait. Since

$$(\hat{\beta_2})^T \hat{\beta_1} = y_2^T X\left(X^T X\right)^{-1}\left(X^T X\right)^{-1} X^T y_1 = \left(X^T X\right)^{-2} y_2^T X X^T y_1 \quad (9)$$

$$(\hat{\beta_1})^T \hat{\beta_2} = y_1^T X\left(X^T X\right)^{-1}\left(X^T X\right)^{-1} X^T y_2 = \left(X^T X\right)^{-2} y_1^T X X^T y_2 \quad (10)$$

$$(\hat{\beta_1})^T = \hat{\beta_1};\ (\hat{\beta_1})^T = \hat{\beta_1} \quad (11)$$

we had

$$\hat{\beta_1}\hat{\beta_2}\left(X^T X\right)^2 = y_2^T X X^T y_1 < n y_2^T y_1 \quad (12)$$

Putting Equation (12) into Equation (5) we got

$$\mathrm{S} > \frac{\left(X^T X\right)^2}{n(n-1)}\begin{bmatrix} \beta_1\beta_1 & \beta_1\beta_2 \\ \beta_2\beta_1 & \beta_2\beta_2 \end{bmatrix} \quad (13)$$

Because $S\mu = \lambda\mu$, where λ was the eigenvalue corresponding to μ, we had

$$\lambda\hat{\beta^*} = \left(X^T X\right)^{-1} X^T [y_1, y_2]\lambda\mu = \left(X^T X\right)^{-1} X^T [y_1, y_2]\mathrm{S}\mu \quad (14)$$

Putting Equation (13) into Equation (5) we got

$$\begin{aligned} \lambda\hat{\beta^*} &> \tfrac{\left(X^T X\right)^2}{n(n-1)}\left(X^T X\right)^{-1} X^T [y_1, y_2]\begin{bmatrix} \beta_1\beta_1 & \beta_1\beta_2 \\ \beta_2\beta_1 & \beta_2\beta_2 \end{bmatrix}\mu \\ &= \tfrac{\left(X^T X\right)^2}{n(n-1)}\left(X^T X\right)^{-1} X^T \left[X\hat{\beta_1} + e_1, X\hat{\beta_2} + e_2\right]\begin{bmatrix} \beta_1\beta_1 & \beta_1\beta_2 \\ \beta_2\beta_1 & \beta_2\beta_2 \end{bmatrix}\mu \end{aligned} \quad (15)$$

By letting $B = [\hat{\beta_1}, \hat{\beta_2}]$ and inserting λ into right-hand side, we got

$$\hat{\beta^*} > \mu\frac{X^T[X,1]\begin{bmatrix} \beta_1 & \beta_2 \\ e_1 & e_2 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}[\beta_1,\beta_2]\left(X^T X\right)}{n(n-1)\lambda} = \frac{X^T X B B^T B\mu\left(X^T X\right)}{n(n-1)\lambda} + \frac{X^T e_1\beta_1 B\mu\left(X^T X\right)}{n(n-1)\lambda} + \frac{X^T e_2\beta_2 B\mu\left(X^T X\right)}{n(n-1)\lambda} \quad (16)$$

The residual error can be considered to be independent of the marker indicator matrix X. $\mathrm{E}(e_1) = 0$ results in $\mathrm{E}(\frac{X^T e_1 \beta_1 B\mu\left(X^T X\right)}{n(n-1)\lambda}) = 0$ and $\mathrm{E}\left(\frac{X^T e_2 \beta_2 B\mu\left(X^T X\right)}{n(n-1)\lambda}\right) = 0$. Provided that the phenotypic correlation coefficient approaches 1, the first eigenvalue can be considered to be

$$\lambda \overset{cor(y_1,y_2)\to 1}{\to} trS = \frac{\left(\beta_1^2 + \beta_2^2\right)\left(X^T X\right)}{n(n-1)}. \tag{17}$$

Therefore, putting Equation (17) into Equation (16), we obtained the β* estimation:

$$\hat{\beta^*} > \frac{X^T X B B^T B\mu\left(X^T X\right)}{n(n-1)\lambda} = \frac{B B^T B\mu}{\beta_1^2 + \beta_2^2} = B\mu = \hat{\beta_1} w_1 + \hat{\beta_2} w_2 \tag{18}$$

where w_1 and w_2 represent elements of the eigenvector μ.

For pleiotropic SNPs, this result indicated that the PCA-based multiple-trait model had a high chi-square statistic for the tested SNP compared to the single-trait model.

2.2. Colocalizing Effect Model

As shown in Figure 1, we assumed that marker 1 had a genuine effect on trait 1, marker 2 had a genuine effect on trait 2, and both were located in the same gene, or within a short distance with a strong linkage disequilibrium (LD). The LD level of the two markers was $r_{LD} = \frac{1}{n} {X_1}^T X_2$, where X_1 and X_2 are the normalized genotypes, with E(X_1) = E(X_2) = 0 and Var (X_1) = Var (X_2) = 1. Similarly, the effects of marker 1 on trait one, marker 2 on trait two, and marker 1 on a pseudo trait are β_1, β_2, and β*, respectively, as in Equations (3)–(5).

Since

$$\left(\hat{\beta_2}\right)^T \hat{\beta_1} = y_2^T X_2 \left({X_2}^T X_2\right)^{-1} \left({X_1}^T X_1\right)^{-1} {X_1}^T y_1 = n^{-2} r^{-1} y_2^T y_1 \tag{19}$$

we had

$$\hat{\beta_1}\hat{\beta_2} nr < {y_1}^T y_2 \tag{20}$$

$$\mathrm{S} > \frac{nr}{n-1} \begin{bmatrix} y_1^T y_1 & y_1^T y_2 \\ y_2^T y_1 & y_2^T y_2 \end{bmatrix} \tag{21}$$

Next, we performed a derivation to estimate β* as in the *single causal variant model*—Equations (13)–(16). Therefore, we had

$$\hat{\beta^*} > \frac{n^2 r B B^T B\mu}{(n-1)\lambda} = r(\hat{\beta_1} w_1 + \hat{\beta_2} w_2) \tag{22}$$

2.3. Simulated Data

We simulated phenotypes based on real data that included 1000 samples and 120,710 SNPs on five chromosomes. The principle of phenotypic simulation is as follows:

$$\mathrm{y} = \sum_i X_i \alpha_i + g + \varepsilon$$

where $g \sim N\left(0, \mathrm{G}\sigma_g^2\right)$ for which σ_g^2 is the additive genetic variance and G is the genomic relationship matrix. α_i is the ith quantitative trait nucleotide (QTN) effect followed by a gamma distribution with a shape parameter of 0.4 and scale parameter 1.66. The polygenic effects vector g was formed by $g = \left(\mathrm{G}^{\frac{1}{2}}\sigma_g\right)^T \tau$, with τ following a normal distribution. The total additive genetic variance can be written as $\sigma_T^2 = \sum \sigma_i^2 + \sigma_g^2$, and the residual error as $\varepsilon \sim N(0, \frac{\left(1-h^2\right)\sigma_T^2}{h^2})$. For the pleiotropic traits simulations, we assumed that each two traits shared 10 common QTNs that contributed 50% of the total genetic variance (σ_T^2).

When simulating low heritability traits, we set the parameters as $h_2 = 0.05$ and $r(e_1,e_2) = 0$. When simulating environmental correlation traits, we set the parameters as $h_2 = 0.5$ and $r(e_1,e_2) = 0.25$.

2.4. Real Data

In the GWAS analysis, a total of 1141 Simmental beef cattle born between 2008 and 2014 composed the experimental population. All cattle were from more than 30 families and were fattened for 8–12 months in a similar environment with the same feed, and slaughtered following the Standard Wholesale Cuts of American Beef guidelines. The phenotypes of three meat cut traits, including the clod weight (CW), fore shank weight (FSW), and heel muscle shank weight (HMSW), were collected during slaughtering. DNA was extracted from the blood samples and genotyped using an Illumina BovineHD BeadChip (Illumina, CA, USA).

Quality control was conducted as follows: (1) Individuals with a call rate < 0.95 and SNPs with a call rate < 0.9 were removed, (2) minor allele frequency < 0.05, and (3) *p*-Value of Hardy–Weinberg equilibrium < 10^{-6}. Finally, a total of 1111 individuals and 608,761 SNPs were left for subsequent analysis. In this study, all phenotypes followed normal distribution and GWAS analyses were implemented using a mixed linear model (MLM). PCA was performed by SAS (Statistical Analysis System) software version 9.4 (SAS Institute Inc., Cary, NC, USA) and genetic parameter estimations were conducted using GCTA (Genome-wide Complex Trait Analysis) [23].

2.5. Power Examination and False Discovery Rate (FDR) Examination

Based on the simulated phenotypes, the power and FDR were calculated under different significant thresholds using a single-trait model and PCA-based multiple-trait model. Power was evaluated as the proportion of QTNs that passed the significance threshold. FDR was defined as the proportion of the non-QTN markers among the identified markers that exceeded the threshold, where the non-QTN markers were markers that were not located 10 Kb upstream or downstream of the QTNs. A total of 100 replicates were conducted for each group, and the average of the 100 replicates was reported.

3. Results

3.1. Simulated Data

We first simulated one set of pleiotropic traits with 10 shared QTNs and $h_2 = 0.5$. Their positions and effect sizes are listed in Table 1. Then, pleiotropic variants were explored using both a single-trait model and a PCA-based multiple-trait model. The $-\log(p)$ and effect standard error (Se Eff) for each QTN are shown in Table 1. Compared with single-trait GWAS, PCA-based multiple-trait GWAS identified additional QTNs. For example, the $-\log(p)$ of the chr1:132347489 locus in PCA-based GWAS was 6.16, and the corresponding values in the two single-trait GWASs was 4.57 and 5.85. If the significant threshold was $p < 10^{-6}$, this locus could be found using PCA-based GWAS, rather than single-trait GWAS.

Table 1. Positions, effects, and *p*-Values of ten quantitative trait nucleotides (QTNs) based on simulated data without environmental correlation.

Chr [a]	Pos (bp)	Trait 1 eff	Trait 2 eff	Single-Trait GWAS				Multiple-Trait GWAS	
				$-\log(p)$ t_1	se eff	$-\log(p)$ t_2	se eff	$-\log(p)$ mt	se eff
1	5167453	1.18	1.66	3.63	0.06	1.87	0.09	3.19	0.01
1	126001364	1.34	1.93	4.38	0.03	3.45	0.04	4.65	0.01
1	128776905	1.83	2.51	1.13	0.13	1.17	0.18	1.33	0.03
1	132347489	1.21	1.91	4.57	0.13	5.85	0.18	6.16	0.03
1	135921964	0.89	1.43	1.73	0.06	4.70	0.08	3.53	0.01
4	28841329	0.93	1.47	1.10	0.04	3.68	0.05	2.54	0.01
4	65810279	1.82	2.38	5.24	0.11	5.22	0.16	6.24	0.02
4	80902019	3.41	5.71	17.55	0.06	30.18	0.08	28.08	0.01
4	115266053	2.20	3.94	10.05	0.06	16.65	0.08	15.70	0.01
5	6270944	0.84	0.94	2.48	0.04	0.87	0.05	1.87	0.01

Note: [a] One of the simulated data results. Pleiotropic traits were simulated based on 10 QTNs. If the significant threshold was a *p*-Value < 10^{-6}, only two QTNs (chr4: 80902019 and chr4: 115266053) could be identified based on single-trait GWAS results. Meanwhile, four QTNs (chr1: 132347489, chr4: 65810279, chr4: 80902019, and chr4: 115266053) could be identified based on PCA-based GWAS results. Shaded QTNs are causal variants only found in PCA-based GWAS. GWAS, Genome-Wide Association Study. Chr, Chromosome. Pos, Position. Eff, effective. Se eff, Standard error of estimated effects.

To facilitate the comparison of the two association strategies, we compared the power and FDR between them in three situations: Medium heritability (h_2 = 0.5), low heritability (h_2 = 0.05), and environmental correlation (h_2 = 0.5, r_e = 0.25). Table 2 shows phenotypic variance and heritability explained by each principal component (PC) in each scenario. The first dimension (PC1) explained more heritability (h_2 = 0.534, 0.052, and 0.580) compared with the second dimension (h_2 = 0.271, 0.035, and 0.130). As shown in Figure 2a, for medium heritability traits, the power of detection of pleiotropic QTNs in PCA-based GWAS was higher than in single-trait GWAS under different significance thresholds. Additionally, the FDR in multiple-trait GWAS was lower than that in single-trait GWAS (Figure 2d). As expected, the power and FDR decreased with the threshold level becoming stringent. For low heritability traits and environmental correlation traits, we obtained similar results (Figure 2b–f). Overall, PCA-based multiple-trait GWAS outperformed single-trait GWAS in the detection of pleiotropic QTNs.

Table 2. Phenotypic variance and heritability explained by each principle component.

Scenario	Heritability	Environmental Correlation	PC1		PC2	
			Phenotypic Variance (SD [a])	Heritability Explained (SD)	Phenotypic Variance (SD)	Heritability Explained (SD)
1	0.5	0	75.98 (25.12)	0.534 (0.04)	14.96 (4.34)	0.271 (0.03)
2	0.05	0	56.78 (17.22)	0.052 (0.01)	39.81 (10.23)	0.035 (0.01)
3	0.5	0.25	89.12 (30.09)	0.580 (0.04)	9.80 (2.11)	0.130 (0.07)

Note: [a] SD: Standard Deviation.

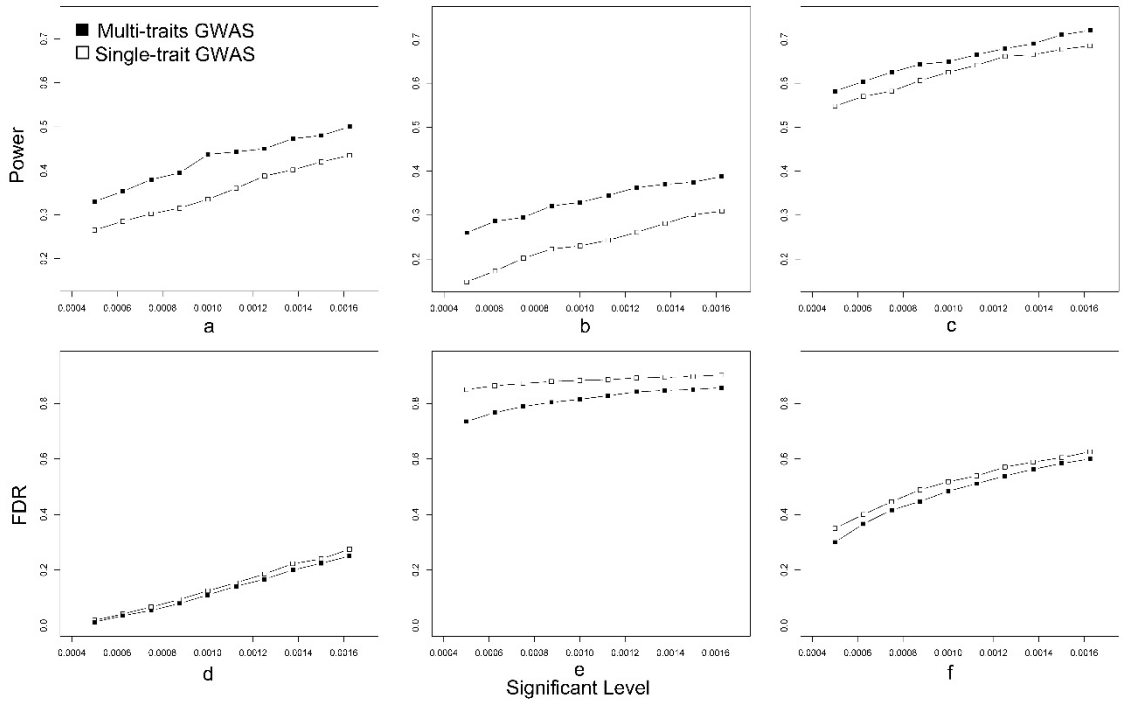

Figure 2. Comparison of power and false discovery rate (FDR) between multiple-trait GWAS and single-trait GWAS. We simulated three situations including medium heritability (**a**,**d**), low heritability (**b**,**e**), and environmental correlation (**c**,**f**). (**a**–**c**) Power under different significant levels. (**d**–**f**) FDR under different significant levels.

For further investigation, we compared the performance of the two models for different minor allele frequencies (MAFs). In each set of simulations, we first randomly simulated pairwise traits by the pleiotropic QTNs regardless of MAF, and then set a significance threshold of the GWAS results (top 0.04% of the total tested SNPs) to define significant SNPs. The power for each SNP was defined as

whether there were significant SNPs harbored by this SNP (1 for harbored, 0 for not harbored). Lastly, based on the power and MAF for each QTN, we fitted trendlines for the two strategies (Figure 3). Overall, PCA-based GWAS outperformed single-trait GWAS. When the MAF of pleiotropic QTNs was less than 0.2, the power difference between them decreased with the reduction of MAF, and when the MAF was greater than 0.2, the differences were maximized and sustained. Since it is hard to define the FDR for each SNP, the relationship between FDR and MAF was not calculated.

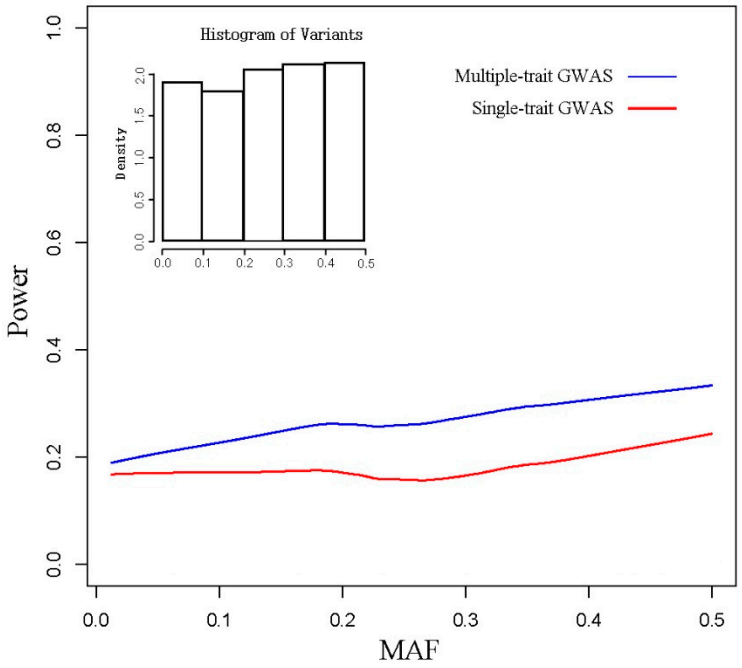

Figure 3. Comparison of detection power between multiple-trait GWAS and single-trait GWAS in different minor allele frequencies. MAF: minor allele frequency. Upper left figure reveals a histogram of the minor allele frequency in the simulated data.

In the colocalizing effect model, to prove Equation (21), we explored the relationship between the capacity of QTL mapping and linkage disequilibrium (LD) of pleiotropic QTNs. Because the value of power/FDR reflects the statistical power of the GWAS model, we found that the capacity of detection was reduced with decreasing LD of pleiotropic QTNs (Figure 4). For pleiotropic QTNs with $r = 0.7$, PCA-based GWAS had a similar power/FDR to single-trait GWAS.

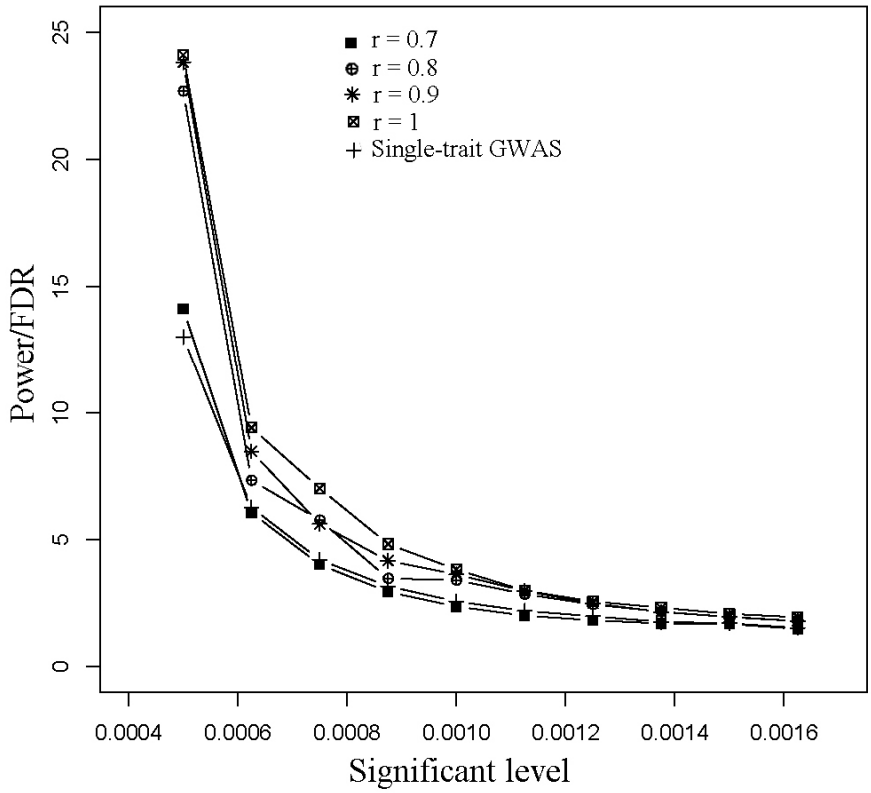

Figure 4. Comparison of power/ False Discover Rate (FDR) in different levels of linkage disequilibrium in the colocalizing effect model.

3.2. Real Data

Three meat cut traits, clod weight (CW), fore shank weight (FSW), and heel muscle shank weight (HMSW), are found in presoma muscles and reflect presoma development in cattle. The heritabilities of the three traits ranged from 0.56 to 0.62, and all three traits had a high phenotypic correlation from 0.76 to 0.82, and genetic correlation from 0.90 to 0.94. The details of the descriptive statistics of the three traits are shown in Table 3.

Table 3. Statistical summary and genetic parameters of three phenotypes.

Trait	Number of Samples	Mean (Kg) (SD)	Heritability	CW	FSW	HMSW
Clod weight (CW)	1111	5.06 (0.88)	0.57	1	0.82 [a]	0.79
Fore shank weight (FSW)	1111	17.03 (3.15)	0.56	0.90 [b]	1	0.76
Heel muscle shank weight (HMSW)	1111	1.07 (0.19)	0.62	0.93	0.94	1

Note: [a] phenotype correlation. [b] genetic correlation.

GWAS analyses for the three traits were conducted using the single-trait GWAS and PCA-based multiple-trait GWAS strategies (Figure 5). The genome-wide significance threshold and suggestive significance threshold were set at 10^{-7} and 10^{-5}, respectively. For CW, only one significant SNP (rs134464739, $p = 3.64 \times 10^{-10}$) was detected on chromosome 4, and no SNPs exceeded the suggestive significance threshold. For FSW, two significant SNPs (rs134464739 and rs134385681, $p > 10^{-5}$), one of which was also identified in CW, were detected on chromosomes 1 and 4, respectively. For HMSW, a total of 24 significant SNPs were found ($10^{-7} > p > 10^{-5}$) on chromosomes 5, 6, and 15. In an approximately 3.5 Mb region (chr6:38550000-42180000), 22 SNPs were associated with the HMSW phenotype, and the most significant SNP was rs137121021, with a p-Value of 1.6×10^{-7}.

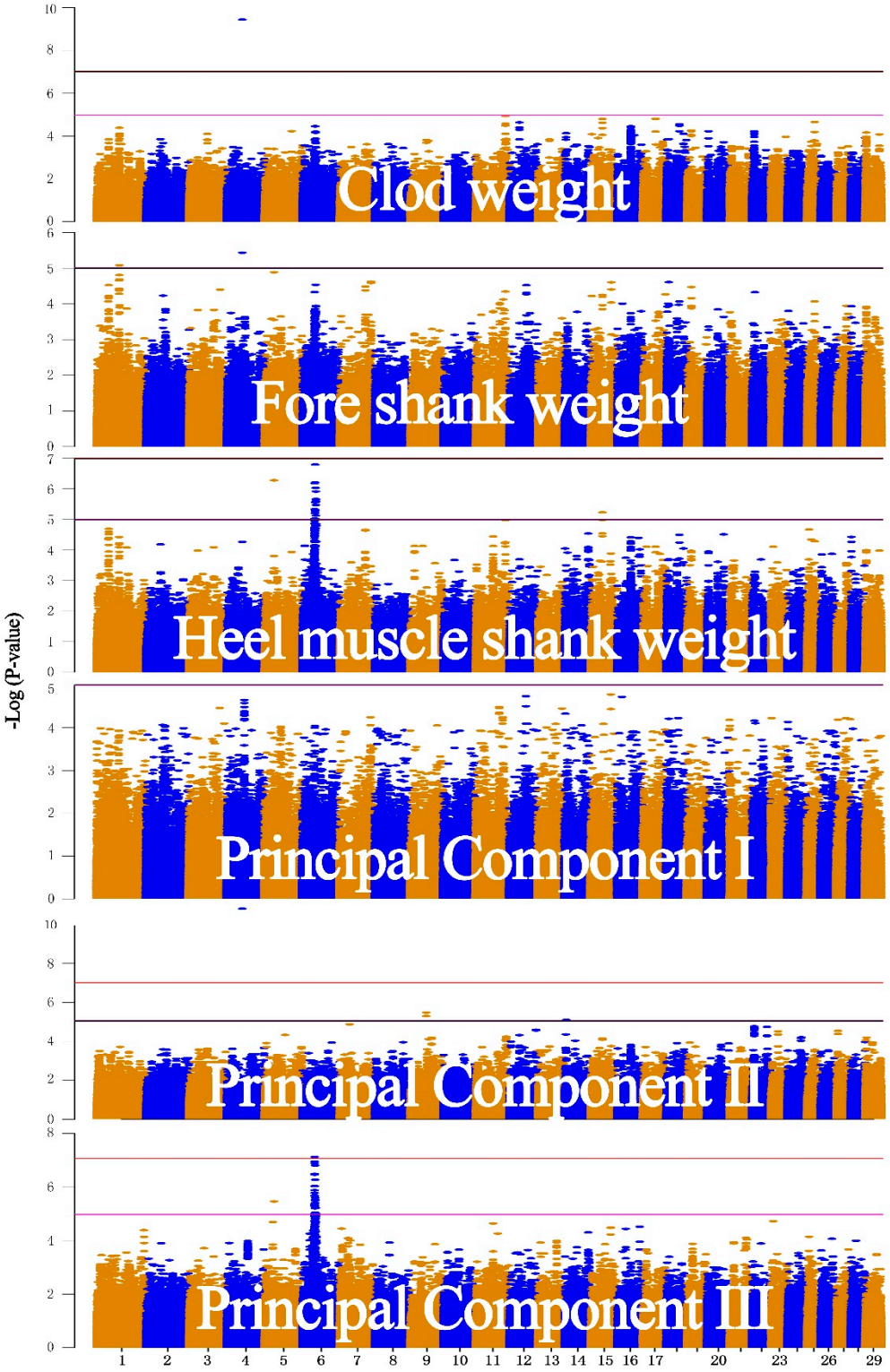

Figure 5. Manhattan plot of the association study results of real cattle data. The three phenotypes are clod weight (CW), fore shank weight (FSW), and heel muscle shank weight (HMSW). The significant level is 10^{-7}, represented by the red line, and the suggestive significant level is 10^{-5}, represented by the pink line.

In the PCA-based GWAS analysis, the three pseudo traits were combined as new phenotypes ($p1$, $p2$, and $p3$), which explained 86.0%, 8.2%, and 5.8% of the total variance, respectively (Table S1). For the

$p1$ GWAS analysis, no significant SNPs were identified. For the $p2$ GWAS analysis, the most significant SNP (rs134464739, $p = 1.39 \times 10^{-11}$) was also found in CW- and FSW-GWASs. Another four associated SNPs, which exceeded the suggestive significance threshold, were located on chromosomes 9 and 14. For the $p3$ GWAS, in the region (chr6:38550000-42180000) where the HMSW trait was associated with 22 SNPs, a total of 31 significant SNPs were found. Another significant SNP, rs134637644 (3.42×10^{-6}), on chromosome 5 was also detected by HMSW. Table S2 lists all significant SNPs identified using both methods.

4. Discussion

The conception of PCA-based QTL mapping was first introduced by Weller in 1996 [13], in which they found canonical variables can represent original traits effectively. Later on, Mangin et al. (1998) [24] proved that multi-trait analysis was more powerful than single-trait analysis for detecting pleiotropic QTL in QTL mapping analysis. In 2008, Lambertus et al. incorporated heritability parameters into a PCA model, which is a powerful association test model. In 2014, Hugues et al. [15] proposed a combined PCA association model that provides greater flexibility and robustness than other PCA methods. In terms of the power of detecting causal SNPs, most multivariate methods, including the PCA-based method, had similar statistical power [25]. In this study, we evaluated potential improvements to this approach using a broad set of data, both synthetic and real. Theoretically, we derived the relationship between multiple-trait GWAS and single-trait GWAS in two pleiotropy models, as shown in Equations (17) and (21). In Equation (17), we assumed $\beta_1 \approx \beta_2$ and $r_{y1,y2} > 0.7$, resulting in $\hat{\beta}^*$ being larger than $\hat{\beta}_1$ and $\hat{\beta}_2$ (Figure S1). We admitted that a simplified general linear model (GLM) might have bias in comparison with a mixed linear model (MLM), and in Equation (16) there should be $\mathrm{cov}(y_1, y_2) \to h_2$ instead of $\mathrm{cov}(y_1, y_2) \to 1$ when the environmental correlation equals 0. However, GLM is approximately equivalent to MLM when analyzing unrelated individuals, and traits with genetic correlation show high phenotypic correlation, indicating that the environmental correlation contributes more. In a pleiotropic trait simulation involving medium heritability, low heritability, and environmental correlation, each pairwise trait shared 10 common QTNs that followed a gamma distribution. We found that multiple-trait GWAS outperformed single-trait GWAS in all three situations, which provides some clues that this approach can be applied to a range of pleiotropic traits. In livestock, detection of pleiotropic QTNs has facilitated the biological understanding of commercial traits, particularly in highly related traits, such as birth weight and weaning weight, as well as milk fat yield and milk protein yield. Additionally, due to taxonomic and binary traits in practical breeding programs, we should further optimize the PCA-based multiple-trait model to combine quantitative traits, taxonomic traits, and binary traits.

For the minor allele frequency (MAF), our results indicated that PCA-based GWAS has significant advantages in pleiotropic QTNs detection when the MAF of QTN is greater than 0.2, while the power improvement gradually reduced when the MAF was less than 0.2. Specifically, for uncommon and rare alleles, the PCA-based strategy had little advantage over the single-trait strategy. In the colocalizing effect model, the estimated effect of a pseudo trait is proportional to the level of linkage disequilibrium (LD) (Equation (21)), and the simulation data supported this view (Figure 4). Under the condition that two traits shared pleiotropic QTNs with $r > 0.7$, PCA-based multiple-trait GWAS was more powerful than single-trait GWAS in detecting QTL regions (Figure 4). Assuming that trait 1 had pleiotropic QTNs with trait 2, it was hard to map to this region using single-trait GWAS because of the low LD between the causal variants and genotyped SNP in the beadchip array. However, when there was a high LD between trait 2's causal variant and the nearby SNP genotyped, this region could potentially be detected in the PCA-based multiple-trait GWAS method after the addition of trait 2.

On the real data, we detected 46 SNPs that were significantly associated with the three traits (Tables S2 and S3). A total of 15 significant SNPs was identified both in single-trait GWAS and multiple-trait GWAS. There were 22 SNPs found only in multiple-trait GWAS, and 9 SNPs found only in single-trait GWAS. Among them, 12 and 18 genes were annotated in multiple-trait GWAS and

single-trait GWAS, respectively, which are growth-related genes or muscle development-related genes, such as *NCAPG* [19,26], *LAP3* [27,28], *KCNIP4* [29], and *LCORL* [26,30]. In contrast, six additional genes were found in single-trait GWAS, including *FBXO45*, *SLIT2*, *SMCO1*, *TCTEX1D2*, *UBXN7*, and *WDR53*, which had not been previously reported in growth-associated studies. Only one additional gene, *MCHR2*, was identified in multiple-trait GWAS. Although single-trait GWAS has annotated more genes, it's result may not be reliable. For example, rs134385681 is a prominent SNP found only in FSW-GWAS which is located in a gene-enriched region, so is likely to be a false positive based on gene annotation. However, *MCHR2* has been reported to be associated with human obesity [31] and a cattle growth trait [32], making it a plausible candidate pleiotropic gene that controls presoma traits.

5. Conclusions

In this study, a PCA-based multiple-trait GWAS model proved to be effective in exploring pleiotropic QTNs in theory and practice. Using this method, we found a plausible candidate gene, *MCHR2*, which is associated with presoma muscle development in cattle.

Supplementary Materials: The following are available online at http://www.mdpi.com/2076-2615/8/12/239/s1, Figure S1: Summation of eigenvectors at different correlation levels based on summated data. Table S1: Component matrix and total variance explained by each principal component. Table S2: Significant SNPs and candidate genes associated with three traits in single-trait GWAS. Table S3: Significant SNPs and candidate genes associated with three traits in PCA-based multiple-trait GWAS.

Author Contributions: J.L. and Y.C. conceived and designed the experiments. W.Z. derived the formulas and wrote the manuscript. Y.C. and X.G. revised the manuscript. X.S. and H.G. performed the analysis. B.Z. and Z.W. collected the experimental database. L.X., L.Z., and X.G. participated in the data collection and dataset analysis. All authors read and approved the final manuscript.

Funding: This work was funded in part by the National Natural Science Foundation of China (31402039, 31372294), National Beef Cattle Industrial Technology System (CARS-37), Chinese Academy of Agricultural Sciences of Technology Innovation Project (CAAS-XTCX2016010, CAAS-ZDXT2018006 and ASTIP-IAS03), the National High Technology Research and Development Program of China (863 Program 2013AA102505-4), and China Scholarship Council (CSC).

Acknowledgments: We are grateful to all scientists and staff of the National Beef Cattle Industrial Technology System in China for supporting the work.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Yang, J.; Benyamin, B.; McEvoy, B.P.; Gordon, S.; Henders, A.K.; Nyholt, D.R.; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **2010**, *42*, 565–569. [CrossRef] [PubMed]
2. Visscher, P.M.; Wray, N.R.; Zhang, Q.; Sklar, P.; McCarthy, M.I.; Brown, M.A.; Yang, J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **2017**, *101*, 5–22. [CrossRef] [PubMed]
3. Solovieff, N.; Cotsapas, C.; Lee, P.H.; Purcell, S.M.; Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **2013**, *14*, 483–495. [CrossRef] [PubMed]
4. Sivakumaran, S.; Agakov, F.; Theodoratou, E.; Prendergast, J.G.; Zgaga, L.; Manolio, T.; Rudan, I.; McKeigue, P.; Wilson, J.F.; Campbell, H. Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **2011**, *89*, 607–618. [CrossRef] [PubMed]
5. Franke, A.; McGovern, D.P.; Barrett, J.C. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **2010**, *42*, 1118–1125. [CrossRef] [PubMed]
6. Iles, M.M.; Law, M.H.; Stacey, S.N. A variant in FTO shows association with melanoma risk not due to BMI. *Nat. Genet.* **2013**, *45*, 428–432. [CrossRef] [PubMed]
7. Teixeira-Pinto, A.; Normand, S.L. Correlated bivariate continuous and binary outcomes: Issues and applications. *Stat. Med.* **2009**, *28*, 1753–1773. [CrossRef]
8. Korte, A.; Vilhjalmsson, B.J.; Segura, A.; Long, Q.; Nordborg, M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **2012**, *44*, 1066–1071. [CrossRef]

9. Zhou, X.; Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **2014**, *11*, 407–409. [CrossRef]
10. Furlotte, N.A.; Eskin, E. Efficient Multiple-Trait Association and Estimation of Genetic Correlation Using the Matrix-Variate Linear Mixed Model. *Genetics* **2015**, *200*, 59–68. [CrossRef] [PubMed]
11. Li, C.; Yang, C.; Gelernter, J.; Zhao, H. Improving genetic risk prediction by leveraging pleiotropy. *Hum. Genet.* **2014**, *133*, 639–650. [CrossRef] [PubMed]
12. Shriner, D. Moving toward System *Genetics* through Multiple Trait Analysis in Genome-Wide Association Studies. *Front. Genet.* **2012**, *3*, 1. [CrossRef] [PubMed]
13. Weller, J.I.; Wiggans, G.R.; Vanraden, P.M.; Ron, M. Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment. *Theor. Appl. Genet.* **1996**, *92*, 998–1002. [CrossRef] [PubMed]
14. Klei, L.; Luca, D.; Devlin, B.; Roeder, K. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet. Epidemiol.* **2008**, *32*, 9–19. [CrossRef] [PubMed]
15. Aschard, H.; Vilhjalmsson, B.J.; Greliche, N.; Morange, P.E.; Tregouet, D.A.; Kraft, P. Maximizing the Power of Principal-Component Analysis of Correlated Phenotypes in Genome-wide Association Studies. *Am. J. Hum. Genet.* **2014**, *94*, 662–676. [CrossRef] [PubMed]
16. Bensen, J.T.; Lange, L.A.; Langefeld, C.D.; Chang, B.L.; Bleecker, E.R.; Meyers, D.A.; Xu, J. Exploring pleiotropy using principal components. *BMC Genet.* **2003**, *4*, S53. [CrossRef] [PubMed]
17. Jiang, L.; Liu, J.; Sun, D.; Ma, P.; Ding, X.; Yu, Y.; Zhang, Q. Genome wide association studies for milk production traits in Chinese Holstein population. *PLoS One* **2010**, *5*, e13661. [CrossRef] [PubMed]
18. 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]
19. Wengang, Z.; Lingyang, X.; Huijiang, G.; Yang, W.; Xue, G.; Lupei, Z.; Bo, Z.; Yuxin, S.; Jinshan, B.; Junya, L.; et al. Detection of candidate genes for growth and carcass traits using genome-wide association strategy in Chinese Simmental beef cattle. *Anim. Prod. Sci.* **2018**, *58*, 224–233.
20. Große-Brinkhaus, C.; Storck, L.C.; Frieden, L.; Neuhoff, C.; Schellander, K.; Looft, C.; Tholen, E. Genome-wide association analyses for boar taint components and testicular traits revealed regions having pleiotropic effects. *BMC Genet.* **2015**, *16*, 36. [CrossRef]
21. Yu, J.; Pressoir, G.; Briggs, W.H.; Vroh, B.I.; Yamasaki, M.; Doebley, J.F.; McMullen, M.D.; Gaut, B.S.; Nielsen, D.M.; Holland, J.B.; et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **2016**, *38*, 203. [CrossRef] [PubMed]
22. Manly, K.F.; Olson, J.M. Overview of QTL mapping software and introduction to map manager QT. *Mamm. Genome.* **1999**, *10*, 327–334. [CrossRef] [PubMed]
23. 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] [PubMed]
24. Mangin, B.; Thoquet, P.; Grimsley, N. Pleiotropic QTL analysis. *Biometrics* **1998**, *54*, 88–99. [CrossRef]
25. Porter, H.F.; O'Reilly, P.F. Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Sci. Rep.* **2017**, *7*, 38837. [CrossRef]
26. Lindholm-Perry, A.K.; Kuehn, L.A.; Oliver, W.T.; Sexten, A.K.; Miles, J.R.; Rempel, L.A.; Cushman, R.A.; Freetly, H.C. Adipose and Muscle Tissue Gene Expression of Two Genes NCAPG and LCORL Located in a Chromosomal Region Associated with Cattle Feed Intake and Gain. *PLoS One* **2013**, *8*, e80882. [CrossRef]
27. Liu, R.; Sun, Y.; Zhao, G.; Wang, F.; Wu, D.; Zheng, M.; Chen, J.; Zhang, L.; Hu, Y.; Wen, J. Genome-Wide Association Study Identifies Loci and Candidate Genes for Body Composition and Meat Quality Traits in Beijing-You Chickens. *PLoS One* **2013**, *8*, e61172. [CrossRef]
28. Xu, L.; Bickhart, D.M.; Cole, J.B.; Schroeder, S.G.; Song, J.; Tassell, C.P.; Sonstegard, T.S.; Liu, G.E. Genomic Signatures Reveal New Evidences for Selection of Important Traits in Domestic Cattle. *Mol. Biol. Evol.* **2015**, *32*, 711–725. [CrossRef]
29. Jin, C.F.; Chen, Y.J.; Yang, Z.Q.; Shi, K.; Chen, C.K. A genome-wide association study of growth trait-related single nucleotide polymorphisms in Chinese Yancheng chickens. *Genet. Mol. Res.* **2015**, *14*, 15783–15792. [CrossRef]

30. Al-Mamun, H.A.; Kwan, P.; Clark, S.A.; Ferdosi, M.H.; Tellam, R.; Gondro, C. Genome-wide association study of body weight in Australian Merino sheep reveals an orthologous region on OAR6 to human and bovine genomic regions affecting height and weight. *Genet. Sel. Evol.* **2015**, *47*, 66. [CrossRef] [PubMed]
31. Meyre, D.; Lecoeur, C.; Delplanque, J.; Francke, S.; Vatin, V.; Durand, E.; Weill, J.; Dina, C.; Froguel, P. A genome-wide scan for childhood obesity-associated traits in French families shows significant linkage on chromosome 6q22.31-q23.2. *Diabetes* **2004**, *53*, 803–811. [CrossRef] [PubMed]
32. Pareek, C.S.; Smoczyński, R.; Kadarmideen, H.N.; Dziuba, P.; Błaszczyk, P.; Sikora, M.; Walendzik, P.; Grzybowski, T.; Pierzchała, M.; Horbańczuk, J.; et al. Single Nucleotide Polymorphism Discovery in Bovine Pituitary Gland Using RNA-Seq Technology. *PLoS One* **2016**, *11*, e0161370. [CrossRef] [PubMed]