G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2015;13(4):146-151
http://dx.doi.org/10.5808/GI.2015.13.4.146

ORIGINAL ARTICLE

Heritability Estimated Using 50K SNPs Indicates Missing Heritability Problem in Holstein Breeding

Donghyun Shin[1], Kyoung-Do Park[2], Sojoeng Ka[1], Heebal Kim[1]*, Kwang-hyeon Cho[3]**

[1]Department of Agricultural Biotechnology, Animal Biotechnology Major, and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul 08826, Korea,
[2]Genomic Informatics Center, Hankyong National University, Anseong 17579, Korea,
[3]Division of Animal Breeding and Genetics, National Institute of Animal Science, Rural Development Administration, Cheonan 31001, Korea

Previous studies in Holstein have shown 35% to 51.8% heritability in milk production traits, such as milk yield, fat, and protein, using pedigree data. Other studies in complex human traits could be captured by common single-nucleotide polymorphisms (SNPs), and their genetic variations, attributed to chromosomes, are in proportion to their length. Using genome-wide estimation and partitioning approaches, we analyzed three quantitative Holstein traits relevant to milk production in Korean Holstein data harvested from 462 individuals genotyped for 54,609 SNPs. For all three traits (milk yield, fat, and protein), we estimated a nominally significant ($p = 0.1$) proportion of variance explained by all SNPs on the Illumina BovineSNP50 Beadchip (h^2_G). These common SNPs explained approximately most of the narrow-sense heritability. Longer genomic regions tended to provide more phenotypic variation information, with a correlation of 0.46∼0.53 between the estimate of variance explained by individual chromosomes and their physical length. These results suggested that polygenicity was ubiquitous for Holstein milk production traits. These results will expand our knowledge on recent animal breeding, such as genomic selection in Holstein.

Keywords: breeding, genomic selection, Holstein, heritability, single-nucleotide polymorphism

Introduction

New-generation sequencing technologies have substantially enhanced the development of genomic tools to assist in breeding decisions. Among many techniques, genotyping technologies have enabled breeders or researchers to identify DNA regions or quantitative trait loci (QTL) associated with a particular phenotypic trait of domesticated animals. Especially, several single-nucleotide polymorphisms (SNPs) and QTL associated with Holstein complex traits, including economic traits related to milk production, were investigated in the approximate 10-year wave of genome-wide association studies (GWASs) [1-3]. In the near past, researchers discovered genetic markers and created marker panels for use in performing marker-assisted selection (MAS). Although MAS provided a first genomic approach to achieve breeding goals to animal breeders, the domesticated animal genome in MAS was taken lightly with narrow perspectives. MAS is restricted when it comes to predicting animal capacity, as it depends on a small number of QTLs that are tagged by markers associated with each trait [4]. Additionally, the gap between the proportion of phenotypic variance accounted for by the top SNPs that reach genome-wide significance level in a GWAS and the heritability estimated from pedigree analyses remained unexplained for most complex traits of human. This was called the missing heritability problem [5], explanations to which have been debated in the field [6].

Recent studies using whole-genome estimation approaches demonstrated that a large proportion of heritability for complex traits of humans can be captured by all common

Received September 9, 2015; Revised December 19, 2015; Accepted December 19, 2015
***Corresponding author:** Tel: +82-2-880-4822, Fax: +82-2-883-8812, E-mail: heebal@snu.ac.kr
****Corresponding author:** Tel: +82-41-580-3362, Fax: +82-41-581-2086, E-mail: ckh1219@korea.kr

SNPs on current genotyping arrays. It implies that there are a large number of variants with an effect that is too small to pass the stringent genome-wide significance level. In animal breeding, genomics selection (GS) has been developed to overcome restriction factors of MAS; it has covered a large number of variants containing small effects. The fundamental difference between MAS and GS as important animal selection tools is that scale GS embraces a much larger number of markers than MAS in predicting animal capacity. GS uses a dense set of markers from across the entire genome to use all markers containing SNPs with small effects. This feature gives GS a profound advantage over MAS.

For predictions based on genomic data for humans, a previous study has already estimated the proportion of phenotypic variance. It was elucidated by the common SNPs all together on a genotyping array for a range of quantitative traits in a large homogenous human sample, using the whole-genome estimation and partitioning approaches [7]. The novelty of our research lies in animal capacity prediction based on genomic data due to the fact that animal evaluation is a core content in the breeding industry. Finally, to conduct our animal breeding research on polygenic inheritance, which was likely to be ubiquitous for complex Holstein traits associated with milk production traits, the genomic selection technique was performed along with various different analyses.

Methods

Korean Holstein data

This study used 462 Holsteins from Korea, and they had information on three phenotypes related to milk production for parity 1 (Fig. 1). Three traits used in this study were milk yield, milk fat, and milk protein; 339 of 462 Holsteins in records of parity 1 had records of parity 2. The individuals produced milk in 63 farms and were born in 2005 to 2012. All candidates were measured for a range of quantitative traits through public surveys for livestock improvement.

Genotyped and imputed data

The genomic DNAs were isolated from snivel by using the nasal collection kit and were genotyped with 54,609 SNPs on the Illumina BovineSNP50 BeadChip. We excluded the SNPs with a missingness rate of 0.05, minor allele frequency (MAF) 0.01, and Hardy-Weinberg equilibrium test p-value 10^{-6} using PLINK [8]. Then, we removed SNPs on the sex chromosome and retained 41,099 autosomal SNPs for further analysis. After this quality control, 462 Holstein genomic data had been imputed using BEAGLE [9].

Estimating and partitioning genetic variance using SNP data

We estimated the genetic relationship matrix (GRM) between all pairs of individuals from all genotyped SNPs For each trait, we then estimated the variance that can be captured by all SNPs using the restricted maximum likelihood approach in a mixed linear model:

$$y = Xb + g_G + e$$

where y is a vector of phenotypes, b is a vector of fixed effects with its incidence matrix X, and g_G is a vector of aggregate effects of all SNPs.

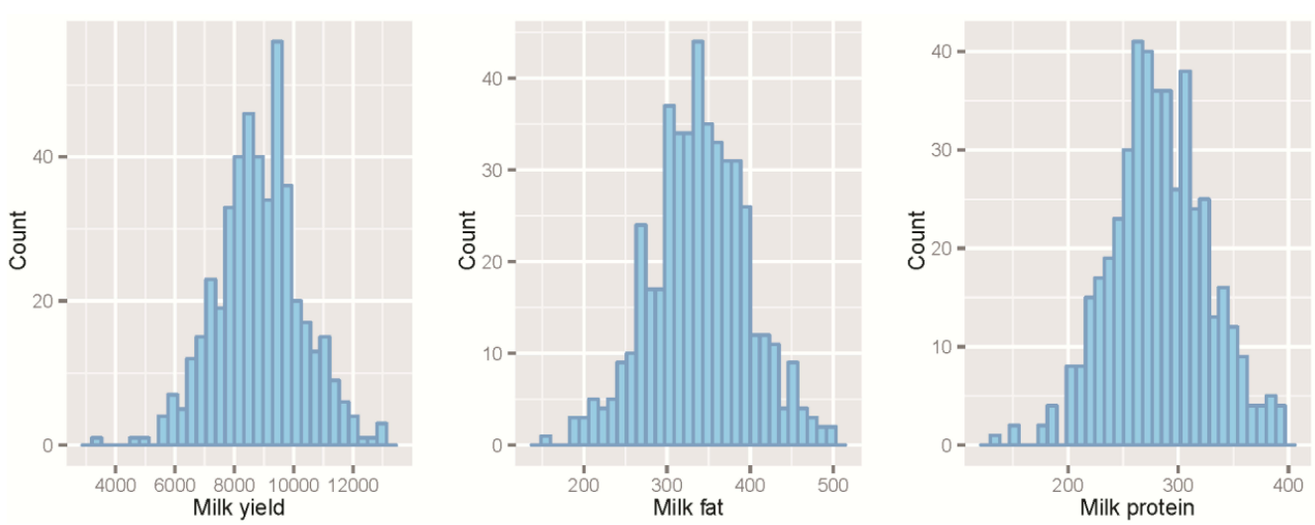

Fig. 1. Phenotype distribution of the three traits related to milk production.

$$\mathrm{Var}(g_G) = A_G \sigma_G^2$$

A_G is the SNP-derived GRM, and σ_G^2 is the additive genetic variance. The proportion of variance explained by all SNPs is defined as

$$h_G^2 = \sigma_G^2 / \sigma_p^2$$

σ_p^2 is the phenotypic variance. Details of the model and parameter estimation have been described elsewhere [10, 11]. When we estimated genetic variance using SNP data, we have to consider seasonal effects (winter, December–February; spring, March–May; summer, June–August; and autumn, September–November based on birth date) as environmental effects, because milk production traits are closely related to season. Then, we performed ANOVA test to investigate the relationship between each trait (parity 1 and 2) and seasonal effect. Milk fat had only a close relation to season in the ANOVA test of both parity 1 ($p = 0.0267$) and 2 ($p = 0.000432$). We counted on seasonal effect in only the analysis including milk fat. In addition, using the same method as above but allowing multiple genetic components to be fit simultaneously in the model, we partitioned h_G^2 into the contributions of genic ($h_{G_g}^2$) and intergenic ($h_{G_i}^2$) regions of the whole genome across all traits. The genic regions were defined as ±0 kb of the 3' and 5' untranslated regions (UTRs). A total of 13,297 SNPs were located within the boundaries of 7001 protein-coding genes for this definition (±0 kb). The total length of the 7001 protein-coding genes was approximately 585 Mb, which covered 21.91% of the genome. We performed these estimating and partitioning genetic variance using SNP data through GCTA [11].

Results

We used the data from the Korea Holstein population. These data were collected from 462 cows recruited from 63 farms in South Korea, genotyped at 54,609 SNPs on the Illumina BovineSNP50 BeadChip. There were 462 individuals and 41,099 autosomal SNPs after quality control (Methods section). All individuals were measured for three traits (milk yield, milk fat, and milk protein), which were related to milk production. The phenotypic correlations between pairwise traits are visualized in Supplementary Fig. 1. Correlations between milk yield and milk protein were much stronger than other pairwise comparisons.

Table 1. Estimates of variance explained by all SNPs for the three traits related to milk production

Trait	Parity	No.	V(G)/Vp (SE)	p-value	Heritability[a]
Milk yield	1	462	0.430 (0.103)	3.41E-06	0.518
	2	339	0.444 (0.129)	2.34E-04	0.431
Milk fat	1	462	0.470 (0.101)	2.06E-08	0.369
	2	339	0.363 (0.137)	1.08E-03	0.35
Milk protein	1	462	0.443 (0.102)	1.04E-06	0.423
	2	339	0.461 (0.128)	7.43E-05	0.392

SNP, single-nucleotide polymorphism.
[a]Estimate of h^2 from pedigree analysis in literature [13].

We estimated the proportion of variance explained by fitting all SNPs in a mixed linear model for each of the three traits. In general, there was a substantial amount of variance explained by all autosomal SNPs after quality control on the Illumina BovineSNP50 BeadChip (σ_G^2) for three traits of the records of two parities, with a mean of 43.51% (a range from 36.3 to 47%) across all three traits of the records of two parities (Table 1). In the records, the estimate of h_G^2 was non-zero and reached the nominal significance level (likelihood ratio test, $p = 0.05$). We compared the estimates of h_G^2 with the narrow-sense heritability h^2, estimated from pedigree analyses in the literature based on Canada Holstein. The value of h_G^2 was similar to h^2 from pedigree analysis in the literature, and all common SNPs explained most (average, 43.51%; range from 36.3% to 47%) of the narrow-sense heritability (average, 41.38%; range from 36.3% to 47%), despite the estimates of h^2 being from different country populations (Table 1). In contrast, when we performed a genome-wide association analysis in the same sample, we tried to identify genome-wide-significant (Bonfferoni correction $p = 0.05$) SNPs for three traits of parity 1. Three SNPs and one SNP were significant in milk yield and fat, respectively, and there were no significant SNPs for milk protein (Supplementary Fig. 2). We hypothesized that there was a major reason for only a few significant large-effect SNPs associated with Holstein traits related to milk production. First of all, many SNPs with small effects affected the three traits related to milk production. Second, according to previous studies, there are many common variants associated with the traits in humans at a nominally significant level, but their effect sizes are too small to be genome-wide-significant [7].

Using the same method as above but allowing multiple genetic components to be fit simultaneously in the model, we then partitioned h_G^2 into the contributions of individual chromosomes for each of the three traits of two parities and plotted the estimate of variance explained by each chromosome (h_C^2) against the chromosome length for each trait. In a previous study, no linear correlation between h_C^2 and chromosome length for any particular traits was observed for most traits in humans as strongly as shown in the other studies. Then, the average value of estimates of h_G^2 over all of tens traits to reduce the sampling error variance, and the

average estimated h_C^2 was strongly correlated with chromosome length. However, we dealt with only three traits related to milk production, and we aimed to identify a linear correlation between h_C^2 and chromosome length for each trait. The squared correlation between h_C^2 and chromosome length using trait information of parity 1 was 0.46, 0.532, and 0.509 for milk yield, fat, and protein, respectively (Fig. 2). The regression slope of the proportion of the genetic variance attributed to each chromosome on the proportion of the genome represented by each chromosome was significant ($p<0.1$), suggesting a proportional relationship between genome length and genetic variance in Holstein. Consistent results were achieved from myriad genetic variants, each possessing a small effect throughout the widespread whole genome in Holstein.

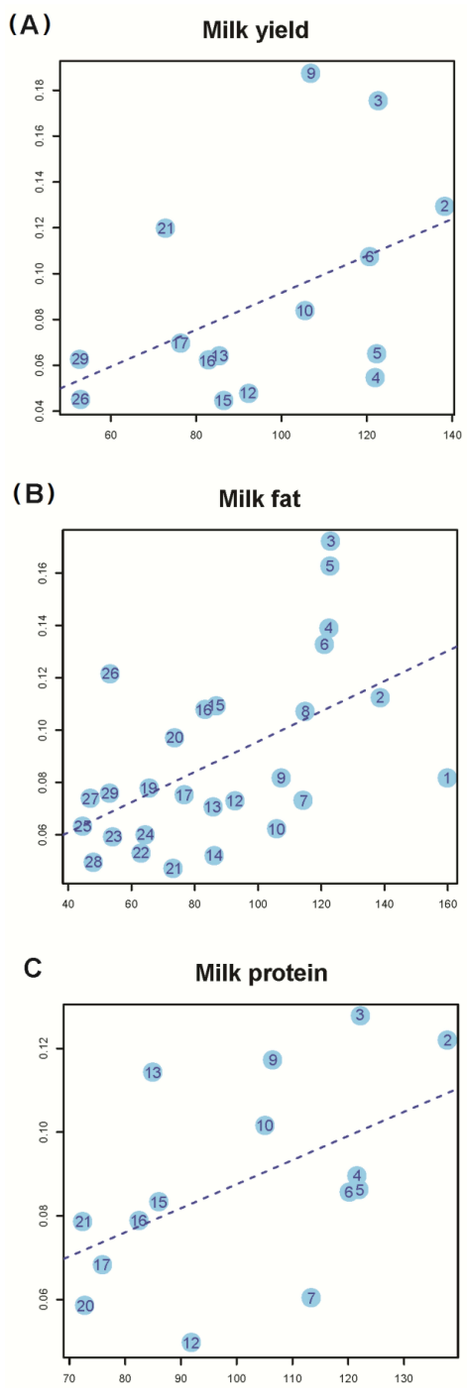

Fig. 2. Proportion of variance attributed to each chromosome of three traits of parity 1 against chromosome length. We show proportion of variance of each chromosome, which is significant based on likelihood test (p-value<0.1). (A) Milk yield (slope = 8.05 E-04, S.E = 4.31 E-04, p-value = 8.46 E-02). (B) Milk fat (slope = 5.79 E-04, S.E = 1.84 E-04, p-value = 4.26 E-03). (C) Milk protein (slope = 5.75 E-04, S.E = 2.97 E-04, p-value = 5.24 E-02).

In addition, we partitioned h_G^2 into the contributions of genic ($h_{G_g}^2$) and intergenic ($h_{G_i}^2$) regions of the whole genome (Methods section). We estimated the variance explained by the genic ($h_{C_g}^2$) and intergenic ($h_{C_i}^2$) regions of each chromosome. The numbers of genic and intergenic SNPs on each chromosome are presented in Supplementary Table 2. We showed that the variance explained by the genic (intergenic) regions on each chromosome is also proportional to the total length of the genic (intergenic) regions (Fig. 3).

Discussion

Previous studies using the whole-genome estimation approach have shown that common SNPs explicate a large proportion of heritability for traits in human [10, 12]. The reason why GWASs in this field have not yet identified all common SNPs that contain information on the amount of variation is mainly that many peculiar effects are too small to pass the stringent genome-wide significance level. Therefore, we hypothesized that inheritance of Korean Holstein milk production traits would consist of a lot of genetic variants with small effects. Finally, we estimated and partitioned the genetic variance that met certain conditions, tagged by all common SNPs for three complex traits related to milk production and showing ubiquitous polygenic inheritance of Korean Holstein. The estimates of h_G^2 for three traits were different from 0 at the significance level ($p<0.1$). The estimate of h_G^2 for milk yield of parity 1 was 43.0% (SE = 10.3%), which was smaller than the estimate from a study in Canada (h_G^2 = 52%) [13]. There could be three possible reasons: (1) There is a difference in the reference population size between Korea and Canada. In fact, Koreans import a lot of Canada Holstein semen, but the present reference population is smaller than Canada. (2) The Korean environment is different from Canada. (3) We can consider only one environmental effect (season) due to data size in this study. The estimate for milk fat (h_G^2 = 47.0%, SE =

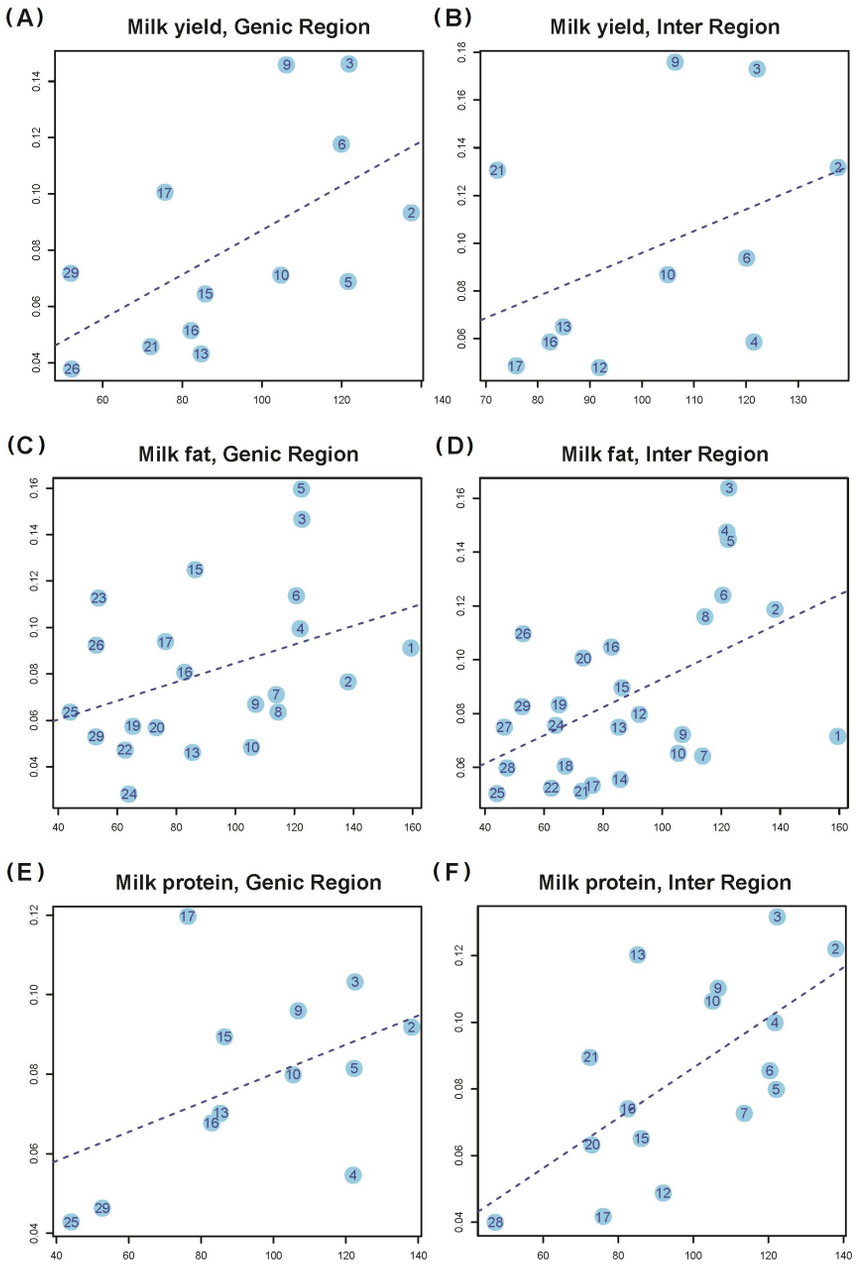

Fig. 3. Estimates of the variance explained by all SNPs in genic (intergenic) regions for 3 traits (a, b, and c indicate milk yield, fat, and protein, respectively) against length of genic (intergenic) DNA. Shown in panels (A), (C), and (E) are the results for the genic SNPs, and shown on panels (B), (D), and (F) are the results for intergenic SNPs, under the assumption that genic regions is Wor0 Kb of UTRs. (A) Milk yield, genic region (slope = 7.89 E-04, S.E = 3.30 E-04, p-value = 3.56 E-02). (B) Milk yield, intergenic region (slope = 9.11 E-04, S.E = 6.70 E-04, p-value = 2.07 E-01). (C) Milk fat, genic region (slope = 4.04 E-04, S.E = 2.20 E-04, p-value = 8.01 E-02). (D) Milk fat, intergenic region (slope = 5.23 E-04, S.E = 1.78 E-04, p-value = 7.02 E-03). (E) Milk protein, genic region (slope = 3.66 E-04, S.E = 2.24 E-04, p-value = 1.34 E-01). (F) Milk protein, intergenic region (slope = 7.53 E-04, S.E = 2.24 E-04, p-value = 7.20 E-03).

10.1%) and protein (h_G^2 = 44.3%, SE = 10.2%) in Korea was larger than in Canada (fat h_G^2 = 36.9%, protein h_G^2 = 42.3%).

It is demonstrated by the genome partitioning analysis that there was a significant linear relationship between the estimates of variance explained by individual chromosomes and chromosome length (Fig. 2). If we define the genic region as ±0 kb of the UTRs, the correlation between variance explained and chromosome length was stronger in the intergenic regions than in the genic regions (Fig. 3). A previous study showed by a number of analyses that the result in this was driven neither by the difference between the number of SNPs in genic regions and in intergenic regions nor by the difference in MAF distribution between genic and intergenic SNPs [7]. If trait-associated genetic variants are enriched in functional elements, such as introns and UTRs, and diluted in exons, the relationship between the explained variance and DNA length will be attenuated in the genic region, as in human. However, it could possibly be just a sampling problem. Because our data size was smaller than our expectancy, some chromosomes did not reach the significance level (p = 0.01) in the genome partitioning analysis. In the genome partitioning analysis for milk yield using SNPs in the intergenic region and milk protein using SNPs in the genic region, the linear regression results were

also insignificant, and we could identify some outliers. Moreover, when these outliers were removed, the linear regression results were significant (E in Fig. 3). We hypothesize that some chromosomes could have a slightly larger effect than our expectation, based on the fitted line in special cases (chromosome 17 or 4 in milk protein).

We showed by whole-genome estimation and partitioning analyses that Holstein milk production traits appeared to be highly polygenic, which means that there were a large number of genetic variants segregating in the population with a small effect widely distributed across the whole genome. All common SNPs expressed most of the heritability on average over all 3 traits analyzed in this study. The slight, remaining unexplained heritability could be due to causal variants, including common and rare ones that are not well tagged by SNPs on this Illumina SNP chip, or possibly because the heritability was over-estimated in the pedigree study. In conclusion, heritability consists of many variants with small effects for Holstein milk production traits in Korea. Overall, this research will become a cornerstone for genomic selection applications in animal breeding.

Supplementary materials

Supplementary data, including two figures and two tables, can be found with this article online at http://www/genominfo.org/src/sm/gni-13-146-s001.pdf.

Acknowledgments

The work was supported by a grant from the Next-Generation Biogreen 21 Program (Project No. PJ01134905), Rural Development Administration, Republic of Korea.

References

1. Jiang L, Liu J, Sun D, Ma P, Ding X, Yu Y, *et al*. Genome wide association studies for milk production traits in Chinese Holstein population. *PLoS One* 2010;5:e13661.
2. Fontanesi L, Calo DG, Galimberti G, Negrini R, Marino R, Nardone A, *et al*. A candidate gene association study for nine economically important traits in Italian Holstein cattle. *Anim Genet* 2014;45:576-580.
3. Meredith BK, Kearney FJ, Finlay EK, Bradley DG, Fahey AG, Berry DP, *et al*. Genome-wide associations for milk production and somatic cell score in Holstein-Friesian cattle in Ireland. *BMC Genet* 2012;13:21.
4. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* 2009;10:381-391.
5. Maher B. Personal genomes: the case of the missing heritability. *Nature* 2008;456:18-21.
6. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;461:747-753.
7. Yang J, Lee T, Kim J, Cho MC, Han BG, Lee JY, *et al*. Ubiquitous polygenicity of human complex traits: genome-wide analysis of 49 traits in Koreans. *PLoS Genet* 2013;9:e1003355.
8. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559-575.
9. Browning BL, Browning SR. A fast, powerful method for detecting identity by descent. *Am J Hum Genet* 2011;88:173-182.
10. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, *et al*. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010;42:565-569.
11. 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.
12. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, *et al*. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* 2011;43:519-525.
13. Miglior F, Sewalem A, Jamrozik J, Bohmanova J, Lefebvre DM, Moore RK. Genetic analysis of milk urea nitrogen and lactose and their relationships with other production traits in Canadian Holstein cattle. *J Dairy Sci* 2007;90:2468-2479.