A single-step genome wide association study on Body Size Traits using imputation-based whole-genome sequence data in Yorkshire pigs

Huatao Liu
China Agricultural University College of Animal Science and Technology
Hailiang Song
China Agricultural University College of Animal Science and Technology
Yifan Jiang
China Agricultural University College of Animal Science and Technology
Yao Jiang
China Agricultural University College of Animal Science and Technology
Fengxia Zhang
China Agricultural University College of Animal Science and Technology
Yibing Liu
China Agricultural University College of Animal Science and Technology
Yong Shi
China Agricultural University College of Animal Science and Technology
Xiangdong Ding
China Agricultural University
Chuduan Wang (cdwang@cau.edu.cn)
China Agricultural University

Research

Keywords: pigs, body size traits, ssGWAS

Posted Date: November 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-101177/v1

A single-step genome wide association study on Body Size Traits using imputation-based whole-genome sequence data in Yorkshire pigs

Huatao Liu, Hailiang Song, Yifan Jiang, Yao Jiang, Fengxia Zhang, Yibing Liu, Yong Shi, Xiangdong Ding*, Chuduan Wang*

National Engineering Laboratory for Animal Breeding, Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing 100193, P.R. China;

Email addresses:

Huatao Liu, 925457179@qq.com

Hailiang Song, 530638625@qq.ocm

Yifan Jiang, jiangyf2020_2020@163.com

Yao Jiang, jiangyao133996@126.com

Fengxia Zhang, 895315294@qq.com

Yibing Liu, 279547013@qq.com

Yong Shi, 2273190789@qq.com

Xiangdong Ding, xding@cau.edu.cn

Chuduan Wang, cdwang@cau.edu.cn

Abstract

Background: The body shape of pig is the most direct production index of pig, which can fully reflect the growth status of pig and is closely related to some important economic traits. In this study, genome-wide association study on seven body size traits, the body length (BL), height (BH), chest circumference (CC), abdominal circumference (AC), cannon bone circumference (CBC), rump width (RW) and chest width (CW) were conducted in Yorkshire pigs.

Methods: Illumina Porcine 80K SNP chip were used to genotype 589 of 5,572 Yorkshire pigs with body size records, and then the chip data was imputed to sequencing data. After quality control of imputed sequencing data, 784,267 SNPs were obtained, and the averaged linkage disequilibrium (r^2) was 0.191. We used the single-trait model and the two-trait model to conduct single-step genome wide association study (ssGWAS) on seven body size traits.

Results: A total of 198 significant SNPS were finally identified according to the P value and the contribution to the genetic variance of individual SNP. 11 candidate genes (CDH13, SIL2, CDC14A, TMRPSS15, TRAPPC9, CTNND2, KDM6B, CHD3, MUC13, MAPK4 and HMGA1) were found to be associated with body size traits in pigs, KDM6B and CHD3 jointly affect AC and CC, and MUC13 jointly affect RW and CW. These genes are involved in the regulation of bone growth and development as well as the absorption of nutrients and are associated with obesity. HMGA1 is proposed as strong candidate gene for body size traits because of its

important function and high consistency with other studies regarding the regulation of body size traits. Our results could provide valuable information for pig breeding based on molecular breeding.

Keywords: pigs, body size traits, ssGWAS

Introduction

Pork is widely used as an important animal protein resource and has become one of the main sources of human protein. Commercial pigs (e.g. Duroc, Yorkshire and Landrace pigs) have the characteristics of fast growth, high feed utilization rate, high lean meat rate and obvious economic benefits. Therefore, it is not only a large number of breeding production, but also the focus of research. The body size trait is one kind of important phenotypic trait that can reflect the overall appearance of animals. Compared with the description of physical appearance, body size traits can objectively reflect the response of pigs to environment and other aspects[35]. In pig breeding, the body shape character index is often used as the most direct production index of pig. Body size is a typical quantitative (or complex) trait, understanding the genetic mechanism of body size differences among individuals can effectively help control the growth and production of animals[34]. At present, there are many researches on genetic parameters of pig external traits, which accelerate the process of genetic improvement of related traits. With the development of molecular

61 biotechnology, many studies have been carried out to clarify the genetic basis of pig
62 body size traits.

63 By far, 1172 QTLs have been found related to body size traits in pigs according to
64 PigQTLdb database (http://www.genome.iastate.edu/cgibin/QTLdb/ss/index).
65 Although a range of researches have been done in QTL mapping, wide confidence
66 intervals (covering more than 20 CM) for the positions of QTL remain that have
67 rarely been replicated[39; 42]. A new research era was initiated with advances in
68 single nucleotide polymorphism (SNP) chip and sequencing technology, and genome
69 wide association study (GWAS) has become one of the most efficient methods to
70 detect genetic variation in livestock[30]. Compared with traditional QTL localization,
71 GWAS has more advantages in mining the intensity of medium-potency variation
72 sites and defining the accuracy of genome segments containing variation sites[19; 26;
73 38; 41]. Although a large number of genome-wide association studies have been
74 carried out in pigs, only few GWAS focused on identifying genes related to external
75 traits. In particular, the investigation on body height, cannon bone circumference,
76 rump width and other important body size traits are still lacking.

77 Marker density is one key factor affecting the efficiency of GWAS as gene mapping
78 mainly relies on the linkage disequilibrium between causal mutation and markers[9].
79 Whole genome sequence data can definitely meet such requirements. In recent years,
80 with the rapid development of the new generation of sequencing technology, the cost

of sequencing has been reduced rapidly, on one hand, a large number of samples and the subsequent processing of sequence data are still time-consuming and costly, limiting its utilization in genetic analysis. On the other hand, genotype imputation provides one efficient tool to improve the marker density of SNP chip based on sequence data. It can accurately predict the genotypes of polymorphic sites not covered by the widely used SNP chip, allowing more genetic loci to be applied to association analysis and improving the possibility of discovering new pathogenic genes [32; 45].In this study, we used imputation-based whole genome sequence data to carry out GWAS on seven body size traits in pigs.

Materials and methods

Ethics statement

The whole recording procedure of ear tissue samples was carried out in strict accordance with the protocol approved by the Institutional Animal Care and Use Committee (IACUC) at the China Agricultural University. The IACUC of the China Agricultural University approved this study (permit number DK996).

Animals and phenotypes

Yorkshire pigs born 2013-2016 from one pig breeding farm in Beijing were collected in this study. Performance test on seven body size traits were carried out at the body weight of about 100 kg for pigs. In total, 5,572 Yorkshire pigs with phenotypic

100 records and pedigree information were selected. The seven body size traits included
101 body length (BL), body height (BH), chest circumference (CC), abdominal
102 circumference (AC), cannon bone circumference (CBC), chest width (CW) and rump
103 width (RW). **Table 1** presents the descriptive statistics of body weight and seven body
104 size traits. There were 4898 records for AC and 5572 records for the other six body
105 size traits and body weight. Normal test showed all the traits followed normal
106 distribution, and the body weight had the largest standard deviation of 12.59 and
107 coefficient of variation of (12.43%), it was used as a covariate considering its
108 influence on the body size traits in further analysis.

109 **Genotype data and imputation**

110 In this study, 589 out of 5572 Yorkshire pigs with body size records were genotyped
111 using the PorcineSNP80 Bead Chip (Illumina, San Diego, CA), which includes
112 68,528 SNPs across the whole pig genome. In order to improve the marker density,
113 the genotyped animals with another 6103 pigs genotyped with PorcineSNP80 [43]
114 were imputed to whole genome sequence data using Beagle 4.1[10]. A wide
115 collection of 289 sequenced pigs all with average sequencing depth of ~25X from 6
116 different pig breeds were used as reference data for imputation and each breed
117 contained 24 to 94 pigs. The composition of reference data and the SNP calling of
118 these individuals were described by Yan et al.[54]. After SNP calling, 46,766,110
119 SNPs were retained as the reference panel for imputation. On average, the genotype
120 concordance rate across all variants was 92%, which is sufficient for further

121 analysis[43]. After imputation, in this study, the following genotype quality control
122 procedure was carried out using the PLINK software (v1.90)[36]. (1) SNPs with
123 minor allele frequency (MAF) lower than 0.01 and deviated from Hardy - Weinberg
124 equilibrium ($P < 10^{-6}$) were excluded and only variants located on autosomes were
125 used for further analysis;(2) the SNP with call rate less than 0.95 were removed;(3)
126 individuals with call rate less than 0.90 were excluded. In addition, in order to
127 decrease the influence of the dependence of adjacent markers on the high false
128 positive of GWAS analysis, the SNP were further pruned, the SNP with linkage
129 disequilibrium (r^2) in slide window of 50 SNPs less than 0.9 were selected. Finally, all
130 the genotyped animals and 784267 SNPs were retained.

131 ## Statistical models

132 **genetic correlation**

133 According to the information of 5,572 pigs in this study, the restricted maximum
134 likelihood method (AI-REML) in DMU v6.0 software[31] was used to estimate the
135 genetic correlations of seven body size traits.
136 The animal model was used to estimate the genetic parameters:

137 $$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Xb} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{t} + \mathbf{e},$$

138 with

139 $$E\begin{Bmatrix} y \\ a \\ t \\ e \end{Bmatrix} = \begin{Bmatrix} Xb \\ 0 \\ 0 \\ 0 \end{Bmatrix}, \quad Var\begin{Bmatrix} a \\ t \\ e \end{Bmatrix} = \begin{Bmatrix} A\sigma_a^2 & 0 & 0 \\ & I\sigma_t^2 & 0 \\ & & I\sigma_e^2 \end{Bmatrix}$$

140 where, **y** is the vector of phenotypic values of each body size trait; **μ** is the population
141 mean; **b** is the fixed effect of herd-year-season; **a** is the vector of additive genetic
142 effects; **t** is the covariate vector of body weight effects; **e** is a vector of residual
143 effects. **X**, $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are incidence matrices associating **b**, **a** and **t with y**,
144 respectively. **A** is the genetic relationship matrix, five generations of pedigree were
145 traced back to construct A, and σ_a^2 is the additive genetic variance. **I** is the identity
146 matrix of appropriate dimension, σ_t^2 is the variance of body weight effect and σ_e^2 is
147 the residual variance.
148 Subsequently, genetic correlations were calculated based on the variance components
149 as follows:

150 $$r_A = \frac{cov(a_1, a_2)}{\sigma_{a_1}\sigma_{a_2}}$$

151 where, r_A is the genetic correlation between trait 1 and trait 2, a_1 and a_2 represent the
152 additive genetic values of trait 1 and trait 2 for same individuals, cov (a1, a2) and σa1,
153 σa2 refer to the genetic covariance of two traits and the genetic standard deviation of
154 trait 1 and trait 2, respectively.

Genome-wide Association Study

In this study, single-step GWAS (ssGWAS) , which can simultaneously use all the SNP information and utilize the ungenotyped animals with phenotypic records[47], was implemented to identify significant SNPs associated with body size traits. Considering the genetic correlations between body size traits, two-trait ssGWAS model was also conducted on traits with high genetic correlations.

Single-trait ssGWAS

Single-trait ssGWAS model was used for three body size traits BL,BH and CBC.

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\gamma}\mathbf{W} + \mathbf{Zg} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of phenotypic values, $\mathbf{b}$ is the vector of fixed effects including herd-year-season-sex, $\mathbf{W}$ is the covariate of body weight, $\mathbf{g}$ is the vector of additive genetic effects, following a normal distribution of $N(\mathbf{0}, \mathbf{H}\sigma_g^2)$, in which $\mathbf{H}$ is the matrix of additive genetic relationships incorporating both pedigree and genomic information, σ_g^2 is the additive genetic variance, $\mathbf{e}$ is the vector of random residuals with distribution of $N(0, \mathbf{I}\sigma_e^2)$, in which $\mathbf{I}$ is the identity and σ_e^2 is the residual variance. $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{Z}$ is the incidence matrix associating $\mathbf{b}$, $\mathbf{w}$, $\mathbf{g}$ with $\mathbf{y}$, respectively.

The genotyped and ungenotyped animals were considered simultaneously based on a H matrix[4]. The inverse of the H matrix was written as follows:

$$\boldsymbol{H^{-1}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G_w^{-1}} - \boldsymbol{A_{22}^{-1}} \end{bmatrix} + \boldsymbol{A^{-1}}$$

where $\boldsymbol{A^{-1}}$ is the inverse of the numerator relationship matrix, $\boldsymbol{A_{22}^{-1}}$ is only the inverse of the pedigree-based relationship matrix for the genotyped animals, and $\boldsymbol{G_w^{-1}}$ is the inverse of the genomic relationship matrix;, G weight markers were obtained by reciprocals of expected marker variance[46].

The SNP effects could be estimated by ssGWAS. The proportion of genetic variance explained by single SNP was calculated as follows:

$$\frac{\boldsymbol{Var}(\boldsymbol{Z_j}\hat{u}_j)}{\boldsymbol{\sigma_a}^{\mathbf{2}}} \times \mathbf{100}\%$$

where $\boldsymbol{\sigma_a}^{\mathbf{2}}$ is the total genetic variance, $\boldsymbol{Z_j}$ is a vector of the gene content of the jth SNP for all animals, and $\hat{u}_j$ is the estimated marker effect of the jth SNP.

Two-trait ssGWAS

According to the genetic correlation estimations, four body size traits with high genetic correlations (CC and AC, RW and CW) were carried out using two-trait ssGWAS model.

$$\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{b_1} \\ \boldsymbol{b_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\gamma_1} \\ \boldsymbol{\gamma_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{W_1} \\ \boldsymbol{W_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z_1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{g_1} \\ \boldsymbol{g_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix},$$

where $\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{bmatrix}$ is the vector of observation values of trait I and II, $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ are the vector of fixed effects of herd-year-season-sex of trait I and II, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are the incidence matrix associating $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ with $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$, $\begin{bmatrix} \boldsymbol{W_1} \\ \boldsymbol{W_2} \end{bmatrix}$ is the vector of

189 covariate of body weight of trait I and II, $\boldsymbol{\gamma_1}$ and $\boldsymbol{\gamma_2}$ are the regression coefficient
190 associating $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$, $\begin{bmatrix} \boldsymbol{g_1} \\ \boldsymbol{g_2} \end{bmatrix}$ is the vector of additive genetic effects of the two
191 traits, following a normal distribution of $N(\mathbf{0}, \mathbf{H}\otimes\mathbf{M})$,where $\mathbf{M}=\begin{bmatrix} \boldsymbol{\sigma^2_{g1}} & \boldsymbol{\sigma^2_{g12}} \\ \boldsymbol{\sigma^2_{g12}} & \boldsymbol{\sigma^2_{g2}} \end{bmatrix}$ is the
192 additive genetic variance and covariance matrix of the two traits, $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ are the
193 incidence matrix associating $\boldsymbol{g_1}$ and $\boldsymbol{g_2}$ with $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$, $\begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix}$ is the vector of
194 random errors with distribution of $N(\mathbf{0}, \mathbf{I}\otimes\mathbf{R})$, where $\mathbf{I}$ is the identity matrix and
195 $R=\begin{bmatrix} \boldsymbol{\sigma^2_{e1}} & \boldsymbol{\sigma^2_{e12}} \\ \boldsymbol{\sigma^2_{e12}} & \boldsymbol{\sigma^2_{e2}} \end{bmatrix}$ is the residual variance and covariance matrix of the two traits.

196 In this study, for both single trait model and two-trait model of ssGWAS, blupf90
197 [5]was implemented to estimate genomic breeding values (GEBV), and afterwards,
198 based on GEBV, SNP effects and P-values were estimated via postGSf90.The P value
199 of each marker was calculated as follows[3]:

200 $$\boldsymbol{P_i = P_t\left(\frac{\hat{u}_i}{\sqrt{\hat{\sigma}_i^2/n}}, n-1\right),}$$

201 where $\boldsymbol{P_i}$ is the distribution function of t distribution, $\hat{u}_i$ is ith SNP effect, $\boldsymbol{\hat{\sigma}_i^2}$ is the
202 genetic variance of ith SNP, n is the number of animals with ith SNP. In addition, the
203 proportion of genetic variance explained by the ith SNP could also be calculated as
204 $\boldsymbol{\hat{\sigma}_i^2/\sigma_g^2}$. Manhattan plots of SNP variance were obtained by the "qqman" R
205 package[14].

206 In order to control false positives, the False Discovery Rate (FDR)[6; 52] method for
207 multiple testing was used as follow:

$$FDR=m*P_{Max}/n$$

where m is the number of times to be tested, n is the number of significant SNPs at assigned FDR level, e.g. 0.05. P_{Max} is the genome-wide significance level empirical P-value of FDR adjusted. Based on the P-values of SNPs obtained by ssGWAS, the empirical P-value of FDR adjusted at the genome-wide significance level of 0.05 was calculated on each trait in this study.

Identification of candidate genes

After identifying significant SNPs by ssGWAS, the genes located in the 50Kb downstream and 50 Kb upstream region of the significant SNPs were determined using BedTools[37] and pig reference gene annotation (http://www.ensembl.org/Sus_scrofa/Info/Index/; Sus scrofa 11.1 genome version). Using the R package bioconductor (http://www.bioconductor.org/) to identify the related pathways and functional annotation. QTLdb (http://www.animalgenome.org/cgi-bin/QTLdb/SS/download?file= gbpSS_11.1) was used to annotate significant SNPs located in previously mapped QTLs in pigs. R package 'Cluster Profiler'[55] was used to carry out Gene Ontology (GO) and Kyoto research on annotated candidate genes Encyclopedia of Genes and Genomes (KEGG) enrichment analysis.

Results

Genetic correlations of body size traits

Table 2 shows the genetic correlations of seven body size traits. The genetic correlations ranged from -0.286 to 0.840 with standard errors ranging from 0.028 to 0.106. Among the seven body size traits, chest circumference (CC) and abdominal circumference (AC), chest width (CW) and rump width (RW) had the higher genetic correlations of 0.747 and 0.840 with standard errors of 0.055 and 0.028, respectively. The genetic correlations between other traits were lower than 0.3, and some traits were almost not genetic correlated with other traits, e.g. body length (BL) had very low genetic correlation of -0.010,0.03, -0.01,0.01 with body height (BH), CC, AC, CW, respectively.

Identification of significant SNPs associated with body size traits

Two criteria of P value and SNP effect were respectively used to determine the SNPs associated with body size traits. As to the P value, after the 0.05 significance level of the whole genome was adjusted, the P_{Max} values of FDR-based multiple tests were 9.26E-06 for BL, 1.08E-05 for BH, 1.02E-05 for CBC, 9.74E-06 for AC, 1.05E-05 for CC, 9.60E-06 for RW, and 1.01E-05 for CW. As shown in Table 3, a total of 88 significant SNPs was identified for seven body size traits. The Manhattan plots of the three traits BL, BH and CBC using the single trait model are shown in **Figure 1**. For BL, a total of 9 significant SNPs reached the genome-wide significance level, totally accounting for 0.0085% of the genetic variance. These significant SNPS were located

247 on SSC1, SSC6, SSC8, SSC13, SSC14, SSC16, and SSC17.The SNP at
248 SSC17:33632497 explained the largest genetic variance (0.0029%). For BH, only 6
249 SNPs were genome-wide significant, accounting for a total of 0.0123% of genetic
250 variance. They were located on SSC3, SSC5, SSC14, and SSC16.The interpretation of
251 scc16: 886074 has the largest genetic variance (0.0082%). For CBC, there were 15
252 significant SNPs at the genome-wide level, which explained 0.0267% of the genetic
253 variance, and the most significant SNPs were closely located on SSC1. For the two
254 pairs of genetic correlated traits using the two-trait model, the Manhattan plots of AC
255 and CC, RW and CW are shown in **Figure 2**. In total, 8, 17, 9, and 24 SNPs were
256 identified associated with AC, CC, RW, and CW, respectively, and these SNPs
257 explained 0.0109%, 0.0242%, 0.0099% and 0.0281% of genetic variances for the
258 corresponding traits. For each trait, the genetic variance explained by a single
259 significant SNP was very small, the largest of which for each trait were 0.0051%
260 (SSC5:15137502) , 0.0067% (SSC4:64552365), 0.0038% (SSC9:2330339) and
261 0.0065% (SCC7:115471416), respectively. Although the genetic correlations existed
262 among seven body size traits, no common significant SNPs were found.

263 Considering the small contribution of above significant SNPs to the genetic variance,
264 the proportion of genetic variance explained by each SNP were also illustrated as
265 shown in **Figure 3** in this study. Top 20 SNPs with the largest genetic variance were
266 selected for each trait(**Table 3**), SNPs for BL were located on SSC17, BH on SSC2,
267 SSC5 and SSC16, CBC on SSC7 and SSC4, SNPs with largest genetic variance for

AC and CC are located on SSC12, those for RW and CW were on SCC6 SCC7, SCC13 and SCC17. For each body size trait, BL, BH, CBC, AC, CC, RW and CW, the top 20 SNPs explained 2.01%, 1.56%, 1.63%, 2.39%, 2.32%, 1.54% and 1.23% of the genetic variance, respectively. Interestingly, the top 20 SNPs for AC and CC were same, RW and CW shared half of the 20 SNPs. In total, 110 SNPs with larger proportion of explanatory genetic variance were retained for further analysis **(Stable 1)**.

Identification of candidate genes

All the significant SNPs identified by the two methods were annotated within the 50 Kb downstream and upstream region with reference to the Sus scrofa 11.1 genome assembly. According to the two methods of SNP significance and explained genetic variance, 88 and 110 SNPs were identified without overlapping, and 64 and 40 genes were found near these SNPs and only two of them were common, respectively **(Table 3 and Stable 1)**. Six and seven genes were found to be related to the corresponding body size traits by the two methods. The biological processes and pathways involved in these genes include calcium channel proteins, lipid metabolism, and cell proliferation.

Discussion

The superiority of imputation-based WGS data

Genotype marker density is one important factor affecting the efficiency of GWAS [9].With the increase of marker density, the linkage disequilibrium between markers and the target trait QTL is increased, it is helpful for QTL detection. In previous studies, the advantages of whole genome sequencing data have been demonstrated[49]. However, its high cost hampered the widely application of sequencing data. Genotype imputation was proved efficiently to impute the SNP chip data to sequencing data with high accuracy[20]. Our results indicated that imputation-based WGS data dramatically improved the power of GWAS, among the significant SNPs identified in this study, only 3 out of the 88 significant SNPs were located in the PorcineSNP80 SNP chip, the remaining 85 loci were identified in the sequencing data. Moreover, among the 110 non-repeating loci screened by interpretation variance, 101 are new loci after imputation, which indicates that imputed WGS data adds a lot of useful information

Increasing marker density could lead to high linkage disequilibrium (LD) to improve the resolution of gene mapping, while it may also be a burden[24]. Too high LD between markers will cause noise and increase false positive[50]. One of the strategies to deal with such dilemma is to pre-select SNP, which can be done via SNP selection to only keep a set of SNPs that are mutually uncorrelated[11; 18]. Therefore, we pruned SNPs according to the genome-wide sequence data to reduce the LD degree between SNPs, and retained the loci in the original 80K chip. In this study, 44003 out of the qualified 50179 SNPs in PorcineSNP80 chip according to the

genotype quality control were retained, and the average linkage disequilibrium of the finally used 784 ,267 SNPs is similar to that of the chip data, the average r^2 was 0.191 and 0.195, respectively. This not only retains the original SNPs but also increases a large number of SNPS, and does not cause the increase of LD.

The advantage of ssGWAS

Single SNP regression model is widely used in GWAS to identify the association of SNP with traits of interest, whereas it usually yields a high false-positive rate due to ignoring the linkage disequilibrium between adjacent SNPs. Wang et al.[47] proposed Single-step GWAS (ssGWAS) that combines all the data (genotype, phenotype and pedigree information) in one step. It can simultaneously utilize all the markers compared with the single-marker regression genome-wide association analysis, resulting in higher power and accuracy[48]. In addition, ssGWAS is able to use sliding windows to simultaneously analyze multiple SNPs to reduce errors[8; 17], Wang et al. (2012) reported that ssGWAS achieved accuracy of 0.81 ± 0.02 using 1500 genotype animals, which was more accurate than single SNP regression model[47]. Moreover, ssGWAS can utilize more individuals, the sample size in this study is not very large, but has a large number of phenotypic data of ungenotyped animals. Compared with traditional GWAS, ssGWAS can make full use of this part of information, expand the sample size to a certain extent, improve the accuracy of SNP effect estimation, and further improve the efficiency of SNP identification.

The determination of significant SNP using P value or SNP effect

Theoretically, the SNPs with smallest p values were supposed to explain relatively high proportion of genetic variance. Likewise, the SNPs with large effects should be significantly associated with the trait of interest. However, our results indicated that the SNPs with smallest P values did not have large effects, there was no overlap between the top 20 SNPs with smallest P values and with largest SNP effects for each trait. Therefore, in order to locate QTLs related to traits more accurately and comprehensively, this study identified significant SNPs from both P value and SNP effect. The proportion of genetic variance explained by most the significant SNPs was small (0.00004%-0.00653%) for all traits, and the maximum genetic variance of all SNPs was also not large (0.0557%-0.1205%), perhaps because too many SNPs were used in the sequencing data in this study, leading to small effect of each related SNP for each trait. It also indicates that SNPs controlling body size traits are widely distributed on the genome, fitting well the infinitesimal model. It was reported that for complex traits such as height, action sites are widely distributed across the entire genome, indicating that almost all genes are involved in the regulation of height[7].

Pleiotropic effects can lead to genetic correlation between traits. From the aspect of P value, no overlap of significant SNPs associated with two genetic related trait pairs AC and CC, RW and CW were detected in this study. However, more common SNPs with largest effects (not statistically significant) were found in each pair of

genetic related traits, e.g. the top 20 SNPs with largest effects for AC and CC were completely overlapped, these SNPs were adjacent to each other and located near SCC12:53132997. Therefore, it is speculated that these SNPs constitute an important QTL and jointly affect AC and CC. Similarly, there may be QTLs associated with RW and CW around SSC6:39553559 and SCC13:135373704. In addition, we took 20 SNPs as a sliding window, and found that the top 20 windows with largest genetic effects respective for AC and CC were overlapped, as well for RW and CW. The above results further reflect 'one factor produces multiple effects', suggesting that highly genetic related traits are probably regulated by the same QTL.

Potential Candidate Genes for BL、BH and CBC

The body length (BL) is an important index to investigate the breeding performance of animals. According to bioinformatics analysis, CDH13 near SCC6:5671575 could be used as a candidate gene affecting body length. CDH13 is a unique cadherin[44] that regulates cell adhesion, signal transduction and cell growth[29], and plays an important role in the formation of tissues and organs[22]. The ingestion and transfer of Ca will affect the bone development of body for a long time[27], therefore, CDH13 has a certain influence on the growth and development of the body. For BH, SIL2 was found to be associated with this trait near SCC2:46827557, Proteomic studies showed that SIL1 elevation alters the expression of proteins including crucial players in neurodegeneration, abnormal expression of SIL1 has an impact on the morphology of

368 the body, which can reduce the body size[28]. CBC reflects the physical quality of the
369 animal, whether it is strong or not. There are three candidate genes associated with
370 CBC, CDC14A, TMPRSS15 and TRAPPC9. CDC14A is widely expressed in
371 eukaryotic cell biology of a special kind of highly conservative dual specificity
372 phosphatase, a variety of studies from yeast to human somatic cells have shown that
373 CDC14 involves extensive roles, including embryonic development and body size[1].
374 TMPRSS15 has an impact on the digestive efficiency of animals, and has been found
375 to be associated with the formation of cholesterol in humans and has been shown to
376 be associated with the development of fat and body weight in mice[51]. TMPRSS15
377 has also a higher variance ranking based on the SNP effect. The gene mutation of
378 transporter particle complex 9 (TRAPPC9), a protein subunit of transporter particle II
379 (TRAPPII), can lead to abnormal embryonic development, abnormal dietary behavior,
380 and is associated with body mass index[2; 33].

381 **Potential Candidate Genes for AC and CC**

382 AC and CC are a pair of highly genetically-related body size traits, which determine
383 the body size of animals and are indicators to fatness and thinness. CTNND2 was
384 closely related to AC according to the P value. Studies have shown that CTNND2
385 participates in the regulation of cell proliferation and affects the body node number of
386 zebrafish[57]. It is found that KDM6B and CHD3 jointly affect AC and CC. KDM
387 subfamily 6 enzymes B (KAM6B) plays an important role in repression of

388 developmental genes[25], and has a regulatory effect on chondrocyte differentiation,
389 thus affecting bone growth and development[15]. CDH3 is a calcium-binding protein
390 that is involved in calcium ion binding and protein binding and is associated with
391 diseases such as malnutrition and developmental malformations. Studies have shown
392 that CHD3 regulates the developmental morphology of zebrafish heart, thereby
393 affecting the abdominal circumference and body shape of zebrafish[12].

394 **Potential Candidate Genes for RW and CW**

395 For RW and CW, MUC13 was detected to affect both RW and CW. MUC13
396 promotes cell proliferation and migration, inhibits apoptosis, and reduces adhesion
397 through a number of signaling pathways[40], and has a certain effect on the
398 absorption of intestinal nutrients, thus affecting the growth and development of bone
399 and the organism. It was found that MAPK4 and HMGA1 affect RW and CW of pigs
400 respectively. MAPK4 is mitogen-activated protein kinase 4, which is involved in the
401 absorption and decomposition of sugars and the formation of fat, so it is related to
402 obesity traits[53]. HMGA1 affects the expression of two IGFBP(insulin-like growth
403 factor binding protein) protein species and plays an important role in cell growth and
404 differentiation[13; 21]. Studies have shown that the deletion of HMGA1 gene results
405 in a significant decrease in the body size of mice[16]. Moreover, a large number of
406 studies have shown that HMGA1 is related to the body size character of pigs. Ji et
407 al.[23] found HMGA1 was a candidate gene affecting body size of pig through

genome-wide association analysis. Zhang et al[56] found that HMGA1 is expressed in pig limb cells and affects the growth and differentiation of chondrocytes. Because of the functional importance of HMGA1 and several studies have shown that it is highly associated with body size traits, it is worth being verified in the future.

Conclusion

In this study, among seven body size traits in pigs, CC and AC, CW and RW were highly genetic correlated with correlation of 0.747 and 0.840, respectively. We implemented ssGWAS to identify SNPs associated with body size traits based on two aspects of P value and the proportion of explanatory genetic variance of SNP. In total, 198 SNPs were identified associated with seven body size traits in Yorkshire, correspondingly, 11 genes were related to body size traits, among which HMGA1 could be worth being validated in further study.

Data availability

The ped and the map are not publicly available because the genotyped animals belong to commercial breeding companies, but are available from the corresponding author on reasonable request.

Conflict of interest

The authors declare that they have no competing interest.

Author contributions

XD and CD conceived and supervised the study. HT, HL and LF helped complete the imputation of the chip data and provide technical guidance. HT, JY, HL, YF collected the samples and recorded the phenotypes. JY and LF extracted the DNA for genotyping. HT, FX, YB and SY contributed to the visualization of data. HT and XD wrote and revised the manuscript. All authors read and approved the manuscript.

Funding

The authors are grateful to all the funding agencies. This work was supported by grants for the National Key Research and Development Project (2019YFE0106800, 2018YFD0501000), China Agriculture Research System (CARS-35), Modern Agriculture Science and Technology Key Project of Hebei Province (19226376D), the National Natural Science Foundation of China (31671327).

References:

1. Research progress in function of cell division cycle protein Cdc14A. Journal of Jilin Univeristy. Medicine edition. 2018;44(1671-587X(2018)44:5<1105:XBFLZQ>2.0.TX;2-55): 1105-8.
2. Abbasi AA, Blaesius K, Hu H, Latif Z, Picker-Minh S, Khan MN et al. Identification of a novel homozygous TRAPPC9 gene mutation causing non-syndromic intellectual disability, speech disorder, and secondary microcephaly. Am J Med Genet B Neuropsychiatr Genet. 2017;174(8): 839-45. doi:10.1002/ajmg.b.32602.

3. Aguilar I, Legarra A, Cardoso F, Masuda Y, Lourenco D, Misztal I. Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. GENETICS SELECTION EVOLUTION. 2019;51(28). doi:10.1186/s12711-019-0469-3.

4. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci. 2010;93(2): 743-52. doi:10.3168/jds.2009-2730.

5. Aguilar I, Misztal I, Legarra A, Tsuruta S. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. JOURNAL OF ANIMAL BREEDING AND GENETICS. 2011;128(6SI): 422-8. doi:10.1111/j.1439-0388.2010.00912.x.

6. BENJAMINI Y, HOCHBERG Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY. 1995;57(1): 289-300. doi:10.1111/j.2517-6161.1995.tb02031.x.

7. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: From polygenic to omnigenic. Cell. 2017;169(7): 1177-86. doi:10.1016/j.cell.2017.05.038.

8. Braz CU, Taylor JF, Bresolin T, Espigolan R, Feitosa F, Carvalheiro R et al. Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle. BMC Genet. 2019;20(1): 8. doi:10.1186/s12863-019-0713-4.

9. Brondum RF, Ma P, Lund MS, Su G. Short communication: Genotype imputation within and across Nordic cattle breeds. J Dairy Sci. 2012;95(11): 6795-800. doi:10.3168/jds.2012-5585.

10. Browning BL, Browning SR. A unified approach to genotype imputation and Haplotype-Phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics. 2009;84(2): 210-23. doi:10.1016/j.ajhg.2009.01.005.

11. Calus M, Vandenplas J. SNPrune: An efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. Genet Sel Evol. 2018;50(1): 34. doi:10.1186/s12711-018-0404-z.

12. Cho SH, Lee CH, Gi E, Yim Y, Koh HJ, Kang K et al. The rice rolled fine striped (RFS) CHD3/Mi-2 chromatin remodeling factor epigenetically regulates genes involved in oxidative stress responses during leaf development. Front Plant Sci. 2018;9(364. doi:10.3389/fpls.2018.00364.

13. Cleynen I, Van de Ven WJ. The HMGA proteins: A myriad of functions (Review). Int J Oncol. 2008;32(2): 289-305.

14. D. Turner S. Qqman: An R package for visualizing GWAS results using Q-Q and manhattan plots. Journal of open source software. 2018;3(25): 731. doi:10.21105/joss.00731.

15. Dai J, Yu D, Wang Y, Chen Y, Sun H, Zhang X et al. Kdm6b regulates cartilage development and homeostasis through anabolic metabolism. Ann Rheum Dis. 2017;76(7): 1295-303. doi:10.1136/annrheumdis-2016-210407.

16. Federico A, Forzati F, Esposito F, Arra C, Palma G, Barbieri A et al. Hmga1/Hmga2

488 double knock-out mice display a "superpygmy" phenotype. BIOLOGY OPEN. 2014;3(5):
489 372-8. doi:10.1242/bio.20146759.
490 17. Guerra FP, Suren H, Holliday J, Richards JH, Fiehn O, Famula R et al. Exome
491 resequencing and GWAS for growth, ecophysiology, and chemical and metabolomic
492 composition of wood of Populus trichocarpa. BMC Genomics. 2019;20(1): 875.
493 doi:10.1186/s12864-019-6160-9.
494 18. Hazelett DJ, Conti DV, Han Y, Al OA, Easton D, Eeles RA et al. Reducing GWAS
495 complexity. Cell Cycle. 2016;15(1): 22-4. doi:10.1080/15384101.2015.1120928.
496 19. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and
497 complex traits. Nature Reviews Genetics. 2005;6(2): 95-108. doi:10.1038/nrg1521.
498 20. Hoze C, Fouilloux MN, Venot E, Guillaume F, Dassonneville R, Fritz S et al. High-density
499 marker imputation accuracy in sixteen French cattle breeds. Genet Sel Evol. 2013;45(33.
500 doi:10.1186/1297-9686-45-33.
501 21. Hristov AC, Cope L, Di Cello F, Reyes MD, Singh M, Hillion JA et al. HMGA1 correlates
502 with advanced tumor grade and decreased survival in pancreatic ductal adenocarcinoma.
503 Modern Pathology. 2010;23(1
504): 98-104. doi:10.1038/modpathol.2009.139.
505 22. Iotzova-Weiss G, Freiberger SN, Johansen P, Kamarachev J, Guenova E, Dziunycz PJ et al.
506 TLR4 as a negative regulator of keratinocyte proliferation. PLoS One. 2017;12(10): e185668.
507 doi:10.1371/journal.pone.0185668.
508 23. Ji J, Zhou L, Guo Y, Huang L, Ma J. Genome-wide association study identifies 22 new loci

509 for body dimension and body weight traits in a White DurocxErhualian F2 intercross
510 population. Asian-Australas J Anim Sci. 2017;30(8): 1066-73. doi:10.5713/ajas.16.0679.
511 24. Joiret M, Mahachie JJ, Gusareva ES, Van Steen K. Confounding of linkage disequilibrium
512 patterns in large scale DNA based gene-gene interaction studies. BioData Min. 2019;12(11.
513 doi:10.1186/s13040-019-0199-7.
514 25. Jones SE, Olsen L, Gajhede M. Structural Basis of Histone Demethylase KDM6B Histone
515 3 Lysine 27 Specificity. Biochemistry. 2018;57(5): 585-92.
516 doi:10.1021/acs.biochem.7b01152.
517 26. Klein RJHJ. Complement factor h polymorphism in Age-Related macular degeneration.
518 Science. 2005;308): 385-9.
519 27. Kovacs CS. Maternal mineral and bone metabolism during pregnancy, lactation, and
520 Post-Weaning recovery. Physiol Rev. 2016;96(2): 449-547.
521 doi:10.1152/physrev.00027.2015.
522 28. Labisch T, Buchkremer S, Phan V, Kollipara L, Gatz C, Lentz C et al. Tracking effects of
523 SIL1 increase: Taking a closer look beyond the consequences of elevated expression level.
524 Mol Neurobiol. 2018;55(3): 2524-46. doi:10.1007/s12035-017-0494-6.
525 29. Liu B, Song J, Luan J, Sun X, Bai J, Wang H et al. Promoter methylation status of tumor
526 suppressor genes and inhibition of expression of DNA methyltransferase 1 in non-small cell
527 lung cancer. Exp Biol Med (Maywood). 2016;241(14): 1531-9.
528 doi:10.1177/1535370216645211.
529 30. Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: Challenges and

prospects. Nature Reviews Genetics. 2009;10(8): 565-77. doi:10.1038/nrg2612.

31. Madsen PAJJ. A user's guide to DMU: A package for analysing multivariate mixed models.. 2013.

32. Marchini J, Howie B. Genotype imputation for genome-wide association studies. NATURE REVIEWS GENETICS. 2010;11(7): 499-511. doi:10.1038/nrg2796.

33. Mbimba T, Hussein NJ, Najeed A, Safadi FF. TRAPPC9: Novel insights into its trafficking and signaling pathways in health and disease (Review). Int J Mol Med. 2018;42(6): 2991-7. doi:10.3892/ijmm.2018.3889.

34. Niu P, Kim S, Choi B, Kim T, Kim J, Kim K. Porcine insulin-like growth factor 1 (IGF1) gene polymorphisms are associated with body size variation. GENES & GENOMICS. 2013;35(4): 523-8. doi:10.1007/s13258-013-0098-0.

35. Ohnishi C, Satoh M. Estimation of genetic parameters for performance and body measurement traits in Duroc pigs selected for average daily gain, loin muscle area, and backfat thickness. Livestock Science. 2018;214(161-6. doi:10.1016/j.livsci.2018.05.022.

36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al. PLINK: A tool set for Whole-Genome association and Population-Based linkage analyses. The American Journal of Human Genetics. 2007;81(3): 559-75. doi:10.1086/519795.

37. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6): 841-2. doi:10.1093/bioinformatics/btq033.

38. Risch NMK. The future of genetic studies of complex human diseases. Epidemiology. 1996;9): 350-4.

551 39. Schreiweis MA, Hester PY, Moody DE. Identification of quantitative trait loci associated
552 with bone traits and body weight in an F2 resource population of chickens. Genetics
553 Selection Evolution. 2005;37(6): 677-98. doi:10.1051/gse:2005023.
554 40. Sheng YH, Triyana S, Wang R, Das I, Gerloff K, Florin TH et al. MUC1 and MUC13
555 differentially regulate epithelial inflammation in response to inflammatory and infectious
556 stimuli. Mucosal Immunol. 2013;6(3): 557-68. doi:10.1038/mi.2012.98.
557 41. Simon-Sanchez J, Schulte C, Bras JM, Sharma M, Gibbs JR, Berg D et al. Genome-wide
558 association study reveals genetic risk underlying Parkinson's disease. Nat Genet. 2009;41(12):
559 1308-12. doi:10.1038/ng.487.
560 42. Soller M, Weigend S, Romanov MN, Dekkers JCM, Lamont SJ. Strategies to assess
561 structural variation in the chicken genome and its associations with biodiversity and
562 biological performance. Poultry Science. 2006;85(12
563): 2061-78. doi:10.1093/ps/85.12.2061.
564 43. Song H, Ye S, Jiang Y, Zhang Z, Zhang Q, Ding X. Using imputation-based whole-genome
565 sequencing data to improve the accuracy of genomic prediction for combined populations in
566 pigs. Genet Sel Evol. 2019;51(1): 58. doi:10.1186/s12711-019-0500-8.
567 44. Takeuchi T, Liang SB, Matsuyoshi N, Zhou S, Miyachi Y, Sonobe H et al. Loss of
568 T-cadherin (CDH13, H-cadherin) expression in cutaneous squamous cell carcinoma. Lab
569 Invest. 2002;82(8): 1023-9. doi:10.1097/01.lab.0000025391.35798.f1.
570 45. van Leeuwen EM, Kanterakis A, Deelen P, Kattenberg MV, Slagboom PE, de Bakker PIW
571 et al. Population-specific genotype imputations using minimac or IMPUTE2. NATURE

PROTOCOLS. 2015;10(9): 1285-96. doi:10.1038/nprot.2015.077.

46. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11): 4414-23. doi:10.3168/jds.2007-0980.

47. Wang H, Misztal I, Aguilar I, Legarra A, Muir WM. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet Res (Camb). 2012;94(2): 73-83. doi:10.1017/S0016672312000274.

48. Wang H, Misztal I, Aguilar I, Legarra A, Fernando RL, Vitezica Z et al. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. Front Genet. 2014;5(134. doi:10.3389/fgene.2014.00134.

49. Wang L, Zhang Y, Zhang T, Zhang L, Yan H, Liu X et al. Genotyping by sequencing reveals a new locus for pig teat number. Anim Genet. 2017;48(4): 470-2. doi:10.1111/age.12547.

50. Wang X, Elston RC, Zhu X. The meaning of interaction. Hum Hered. 2010;70(4): 269-77. doi:10.1159/000321967.

51. Wang Z, Xu S, Du K, Huang F, Chen Z, Zhou K et al. Evolution of digestive enzymes and RNASE1 provides insights into dietary switch of cetaceans. Mol Biol Evol. 2016;33(12): 3144-57. doi:10.1093/molbev/msw191.

52. Weller JI, Song JZ, Heyen DW, Lewin HA, Ron M. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. GENETICS. 1998;150(4): 1699-706.

53. Wu YS, Chen YT, Bao YT, Li ZM, Zhou XJ, He JN et al. Identification and verification of potential therapeutic target genes in Berberine-Treated zucker diabetic fatty rats through bioinformatics analysis. PLoS One. 2016;11(11): e166378. doi:10.1371/journal.pone.0166378.

54. Yan G, Qiao R, Zhang F, Xin W, Xiao S, Huang T et al. Imputation-Based Whole-Genome sequence association study rediscovered the missing QTL for lumbar number in sutai pigs. Sci Rep. 2017;7(1): 615. doi:10.1038/s41598-017-00729-0.

55. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R package for comparing biological themes among gene clusters. OMICS. 2012;16(5): 284-7. doi:10.1089/omi.2011.0118.

56. Zhang LC, Li N, Liu X, Liang J, Yan H, Zhao KB et al. A genome-wide association study of limb bone length using a Large White x Minzhu intercross population. Genet Sel Evol. 2014;46(56. doi:10.1186/s12711-014-0056-6.

57. Zhang Y, Zhang R, Ding X, Ai K. EFNB2 acts as the target of miR-557 to facilitate cell proliferation, migration and invasion in pancreatic ductal adenocarcinoma by bioinformatics analysis and verification. Am J Transl Res. 2018;10(11): 3514-28.

Figures

Figure 1. Manhattan plot of the genome-wide association study on three body size traits by using single-trait model ssGWAS. BL, Body length; BH, Body height; CBC, Cannon bone circumference. In the Manhattan plots, negative log10 P-values of the quantified SNPs were plotted against their genomic positions. The x-axis represents the chromosomes, and the y-axis represents the observed -log10(P-value). Different colors indicate various chromosomes. Each trait has a significant threshold of FDR adjusted, for (A) BL, it was 9.26×10^{-6}. Similarly, (B) BH was 1.08×10^{-5}, and (D) CBC was 1.02×10^{-5}.

Figure 2. Manhattan plot of the genome-wide association study on four body size traits by using two-trait model ssGWAS.
AC, Abdominal circumference; CC, Chest circumference; RW, Rump width; CW, Chest width. AC and CC are a pair of traits, RW and CW are a pair of traits.
In the Manhattan plots, negative log10 P-values of the quantified SNPs were plotted against their genomic positions. The x-axis represents the chromosomes, and the y-axis represents the observed -log10(P-value). Different colors indicate various chromosomes. Each trait has a significant threshold of FDR adjusted, for (A) AC, it was 9.74×10^{-6}. Similarly, (B) CC was 1.05×10^{-5}, (C) RW was 9.60×10^{-6}, and (D) CW was 1.01×10^{-5}.

Figure 3. Manhattan plot of the genome-wide association study on seven body size traits and Venn plot of SNPs according to the contribution of SNP to genetic variance by using ssGWAS.
BL, Body length; BH, Body height; CBC, Cannon bone circumference; AC, Abdominal circumference; CC, Chest circumference; RW, Rump width; CW, Chest width. BL, BH and CBC were single-trait models， AC, CC, RW and CW were two-trait models. AC and CC are a pair of traits, RW and CW are a pair of traits.
In the Manhattan plots(A-G), the proportion of genetic variance of the quantified SNPs were plotted against their genomic positions. The x-axis represents the chromosomes, and the y-axis represents the percentage of SNP explaining the genetic variance. Different colors indicate different chromosomes.
Venn plot(H) of SNPs for the two pairs of body size traits, AC and CC, RW and CW are a pair of traits, respectively.

Tables

Table 1 Descriptive statistics for body weight and seven body size traits

Trait[1]	N-obs[2]	Mean	S.D.	CV(%)	Min value	Max value
BL(cm)	5573	108.89	6.18	5.67	88	134
BH(cm)	5573	62.87	2.92	4.64	51	75
CC(cm)	5573	104.58	5.75	5.50	85	126
AC(cm)	4898	113.52	6.31	5.56	94	137
CW(cm)	5572	29.75	2.31	7.76	19	38
RW(cm)	5573	31.64	2.13	6.73	22	40
CBC(cm)	5573	17.98	1.03	5.73	13	23
BW(kg)	5573	101.31	12.59	12.43	61	150

Note: [1]BL=body length, BH= body height, CC=chest circumference, AC=abdominal circumference, CBC= cannon bone circumference, RW= rump width, CW=chest width , [2]N-obs = number of observations

Table 2 Genetic correlations between seven body size traits

Trait[1]	BL	BH	CC	AC	CW	RW	CBC
BL		-0.010(0.088)	0.033(0.092)	-0.014(0.092)	0.014(0.078)	-0.286(0.078)	0.206(0.078)
BH			0.171(0.104)	0.071(0.106)	-0.221(0.091)	-0.217(0.090)	-0.105(0.096)
CC				**0.747(0.055)**	0.255(0.093)	0.127(0.095)	0.197(0.096)
AC					0.153(0.096)	0.204(0.095)	0.202(0.096)
CW						**0.840(0.028)**	0.015(0.086)
RW							-0.032(0.085)
CBC							

Note: [1]BL=body length, BH= body height, CC=chest circumference, AC=abdominal circumference, CBC= cannon bone circumference, RW= rump width, CW=chest width, SE of estimates are in parentheses

651 **Table 3. Significant SNPs and associated genes for seven body size traits**

Trait[1]	Chromosome	Position (bp)	P_value	SNP effect (%)	Gene	Distance	Gene function
BL	6	5671575	2.35E-07	0.00186	CDH13	+13217	cadherin 13
	1	6435744	1.5E-06	0.00017	NA		NA
	1	6472959	1.5E-06	0.00080	PRKN	+38323	parkin RBR E3 ubiquitin protein ligase
	17	33632497	2.62E-06	0.00289	ENSSSCG00000028461	-47405	signal regulatory protein alpha
	13	25520933	4.45E-06	0.00004	ULK4	-8396	unc-51 like kinase 4
	16	1276330	4.57E-06	0.00031	NA		NA
	14	137476010	6.47E-06	0.00054	NA		NA
	8	28933773	7.46E-06	0.00144	NWD2	-23316	NACHT and WD repeat domain containing 2
	13	166328893	8.39E-06	0.00039	NA		NA
BH	16	886074	2.84E-06	0.00817	CTNND2	+28239	alpha-2-macroglobulin like 1
	8	7942460	3.01E-06	0.00083	NA		NA
	3	26586077	4.62E-06	0.00117	CLEC19A	-45911	C-type lectin domain containing 19A
	5	62690928	6.5E-06	0.00004	A2ML1	-42827	alpha-2-macroglobulin like 1
	4	128701315	7.54E-06	0.00152	NA		NA
	14	33580513	9.85E-06	0.00060	HSPB8	+45615	heat shock protein family B (small) member 8
CBC	4	117759672	2.16E-07	0.00279	CDC14A	-34935	cell division cycle 14A
	13	182971424	1.83E-06	0.00420	TMPRSS15	-29625	transmembrane serine protease 15
	17	12868538	1.85E-06	0.00635	PSD3	-43049	pleckstrin and Sec7 domain containing 3
	1	1201299	2.3E-06	0.00025	ENSSSCG00000041157	-47914	NA
	1	1205821	2.3E-06	0.00018	ENSSSCG00000050693	-42855	NA
	1	1220233	2.3E-06	0.00039	ENSSSCG00000045916	-18409	NA
	1	1367723	2.3E-06	0.00064	ENSSSCG00000043714	+5537	NA
	18	21663467	0.000003	0.00400	GRM8	-14659	glutamate metabotropic receptor 8
	14	9698552	3.19E-06	0.00026	ENSSSCG00000049499	9436	NA
	5	7020488	3.46E-06	0.00023	PMM1	-49963	phosphomannomutase 1
	12	50490164	4.08E-06	0.00233	SPNS3	-47230	sphingolipid transporter 3 (putative)
	3	12869355	4.1E-06	0.00259	ENSSSCG00000036217	+18272	NA
	4	10221008	5.38E-06	0.00076	ASAP1	-37722	ArfGAP with SH3 domain, ankyrin repeat an

	2	124456560	5.76E-06	0.00036	PRR16	+6055	proline rich 16
	1	13806583	7.02E-06	0.00142	ENSSSCG00000004081	-2527	NA

652

Trait[1]	Chromosome	Position (bp)	P_value	SNP effect (%)	Gene	Distance	Gene function
AC	8	3249196	1.88E-06	0.00134	AFAP1	-46660	actin filament associated protein 1
	9	14578071	2.31E-06	0.00128	NA		NA
	14	13670622	2.71E-06	0.00048	PRSS55	-4581	serine protease 55
	5	15137502	2.96E-06	0.00513	RHEBL1	-39837	RHEB like 1
	4	5362087	4.48E-06	0.00139	ENSSSCG00000044937	+36176	NA
	7	26363076	5.4E-06	0.00093	NA		NA
	14	43227411	5.77E-06	0.00024	ENSSSCG00000033385	-49062	KIAA1671 ortholog
	16	522752	6.96E-06	0.00014	CTNND2	-3796	catenin delta 2
CC	3	63528527	1.32E-07	0.00015	ENSSSCG00000008250	-41861	catenin alpha 2
	6	19429624	3.27E-07	0.00022	Metazoa_SRP	-49801	Metazoan signal recognition particle
	1	3149903	7.78E-07	0.00047	PDE10A	-28771	phosphodiesterase 10A
	6	120477523	1.95E-06	0.00173	FHOD3	-40874	formin homology 2 domain containi
	10	56219300	2.01E-06	0.00203	ITGB1	-46293	integrin subunit beta 1
	17	18990746	2.63E-06	0.00004	ANKEF1	-32351	ankyrin repeat and EF-hand domain
	17	18997949	2.63E-06	0.00008	ANKEF1	-37701	ankyrin repeat and EF-hand domain
	2	122228151	3.03E-06	0.00105	ENSSSCG00000051343	-14167	NA
	2	122235537	3.03E-06	0.00074	ENSSSCG00000051343	-21553	NA
	16	33630686	3.58E-06	0.00018	NA		NA
	16	33638300	3.58E-06	0.00079	NA		NA
	16	5533970	4.64E-06	0.00394	ENSSSCG00000016791	+16579	NA
	12	5297390	5.41E-06	0.00092	RNF157	-48707	ring finger protein 157
	4	64552365	5.7E-06	0.00666	ENSSSCG00000042029	-24706	NA
	10	43341283	7.22E-06	0.00064	CUBN	-39687	cubilin
	8	21799389	8.84E-06	0.00030	ENSSSCG00000050984	-18261	NA

	10	60737384	9.42E-06	0.00449	ENSSSCG00000011121	-24200	CUGBP Elav-like family member 2
RW	8	137165913	5.64E-07	0.00049	NA		NA
	9	2330339	2.74E-06	0.00381	SYT9	-11533	synaptotagmin 9
	1	38033383	4.19E-06	0.00149	NKAIN2	-12912	sodium/potassium transporting ATP
	3	63682227	5.03E-06	0.00010	NA		NA
	11	32555905	6.78E-06	0.00015	DIAPH3	+46764	diaphanous related formin 3
	16	48600234	7.1E-06	0.00118	ENSSSCG00000046085	-23005	NA
	16	48696355	7.1E-06	0.00155	ENSSSCG00000039883	+49947	NA
	1	100210738	7.97E-06	0.00095	MAPK4	+8278	mitogen-activated protein kinase 4
	1	100335688	7.97E-06	0.00017	MAPK4	-49706	mitogen-activated protein kinase 4
CW	8	132277288	8.17E-07	0.00009	PTPN13	-27877	protein tyrosine phosphatase non-re
					MAPK10	-27281	mitogen-activated protein kinase 10
	7	115471416	9.52E-07	0.00653	PPP4R4	-18233	protein phosphatase 4 regulatory sul
	14	37118119	1.09E-06	0.00289	ENSSSCG00000051786	-2275	NA
	14	37165658	1.09E-06	0.00051	ENSSSCG00000051786	-49874	NA
	14	37230969	1.09E-06	0.00039	ENSSSCG00000051786	-9755	NA
	2	80016213	2.07E-06	0.00192	COL23A1	-46865	collagen type XXIII alpha 1 chain
	14	139878474	2.34E-06	0.00437	TCERG1L	-46513	transcription elongation regulator 1
	12	49725382	2.67E-06	0.00112	TRPV1	-33424	transient receptor potential cation ch
	16	35012960	3.46E-06	0.00029	DDX4	-46102	DEAD-box helicase 4
	7	115132809	4.65E-06	0.00069	ENSSSCG00000002464	-31787	proline rich membrane anchor 1
	16	73800572	4.82E-06	0.00153	U6	+41611	U6 spliceosomal RNA
	16	73812833	4.82E-06	0.00172	U6	+29350	U6 spliceosomal RNA
	16	73816240	4.82E-06	0.00175	U6	+25343	U6 spliceosomal RNA
	9	107845695	5.74E-06	0.00014	ENSSSCG00000032905	-7136	NA
	8	76456715	6.42E-06	0.00057	ENSSSCG00000042273	-45304	NA
	3	131731345	6.59E-06	0.00001	ENSSSCG00000049751	-23407	NA
	3	131738702	6.59E-06	0.00016	ENSSSCG0000004975	-30764	NA

				1		
3	131744661	6.59E-06	0.00016	ENSSSCG00000049751	-36723	NA
3	131756951	6.59E-06	0.00025	ENSSSCG00000049751	-43896	NA
3	131758601	6.59E-06	0.00024	ENSSSCG00000049751	-45546	NA
5	61521505	7.76E-06	0.00033	ENSSSCG00000033403	-14845	C-type lectin domain family 7 meml
16	67601687	8.2E-06	0.00052	ENSSSCG00000049229	-42425	NA
7	92897102	4.65E-06	0.00109	HMGA1	+25885	high mobility group AT-hook 1
15	11796106	9.96E-06	0.00074	NA		NA

653 Note: [1]BL=body length, BH= body height, CC=chest circumference, AC=abdominal
654 circumference, CBC= cannon bone circumference, RW= rump width, CW=chest
655 width, gene effect= proportion of genetic variance explained

656 **Table 4 Overview of ssGWAS location for the percentage that explains the**
657 **proportion of genetic variance**

Trait[1]	20 SNPs distributions of maximum effect	SNPs of the maximum effect	Top 20 SNPs effect (%)	Number of nearest gene	Candidate gene
BL	SSC17	17_7477978	0.117	4	
BH	SSC2、SSC5、SSC16	2_46827557	0.08.7	8	SIL1
CBC	SSC7、SSC4	7_55099416	0.101	17	TRAPPC9
AC	SSC12	12_53181656	0.128.	8	KDM6B CHD3
CC	SSC12	12_53169477	0.129	8	KDM6B CHD3
RW	SSC6、SSC7、SSC12、SSC13、SSC17	17_13172524	0.099	15	MUC13
CW	SSC6、SSC7、SSC13、SSC17	6_39554872	0.070	28	MUC13

658 Note: [1]BL=body length, BH= body height, CC=chest circumference, AC=abdominal
659 circumference, CBC= cannon bone circumference, RW= rump width, CW=chest
660 width, gene effect= proportion of genetic variance explained
661

662 # Additional files

Additional file Table S1

Format: DOCX

Title: Each trait explains the 20 SNPs with the greatest genetic variance.

Figures

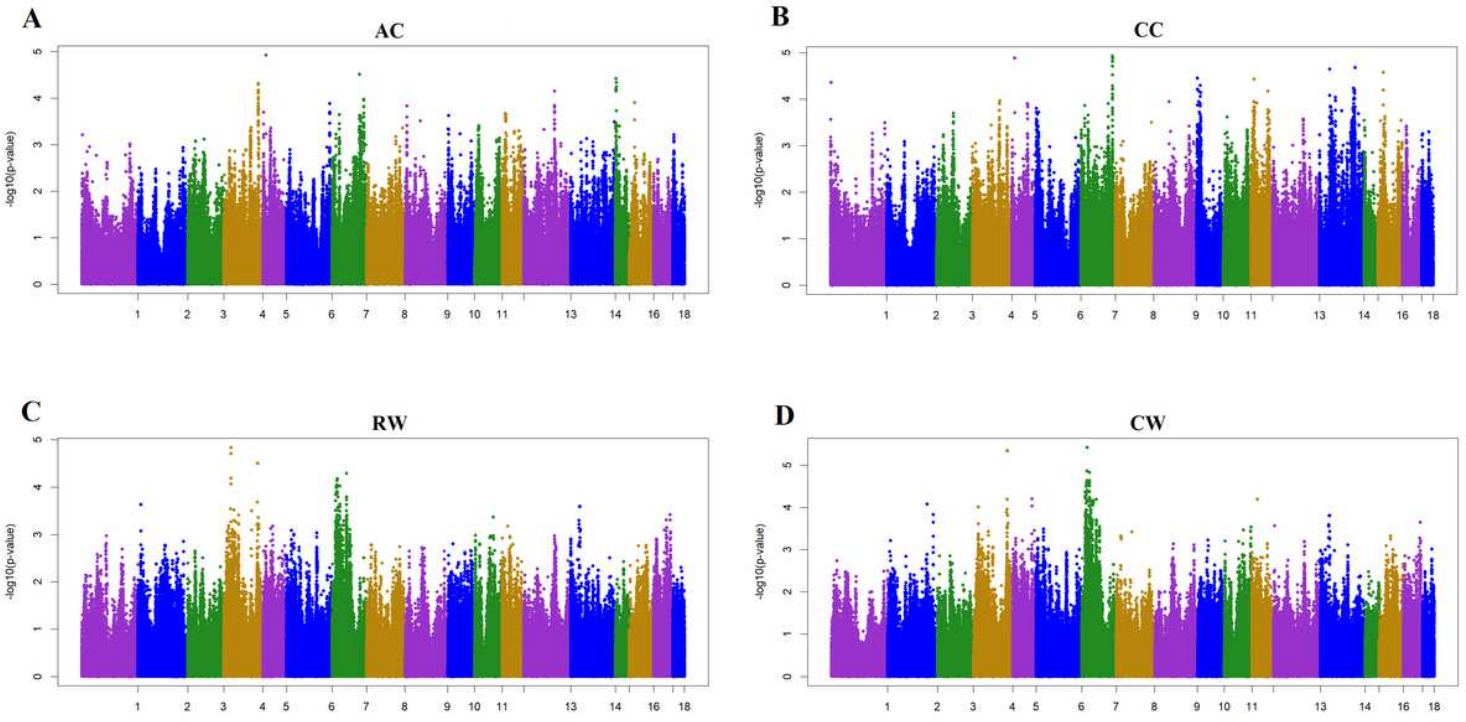

Figure 1

Manhattan plot of the genome-wide association study on three body size traits by using single-trait model ssGWAS. BL, Body length; BH, Body height; CBC, Cannon bone circumference. In the Manhattan plots, negative log10 P-values of the quantified SNPs were plotted against their genomic positions. The x-axis represents the chromosomes, and the y-axis represents the observed -log10(P-value). Different colors indicate various chromosomes. Each trait has a significant threshold of FDR adjusted, for (A) BL, it was 9.26 × 10−6. Similarly, (B) BH was 1.08 × 10−5, and (D) CBC was 1.02 × 10-5.

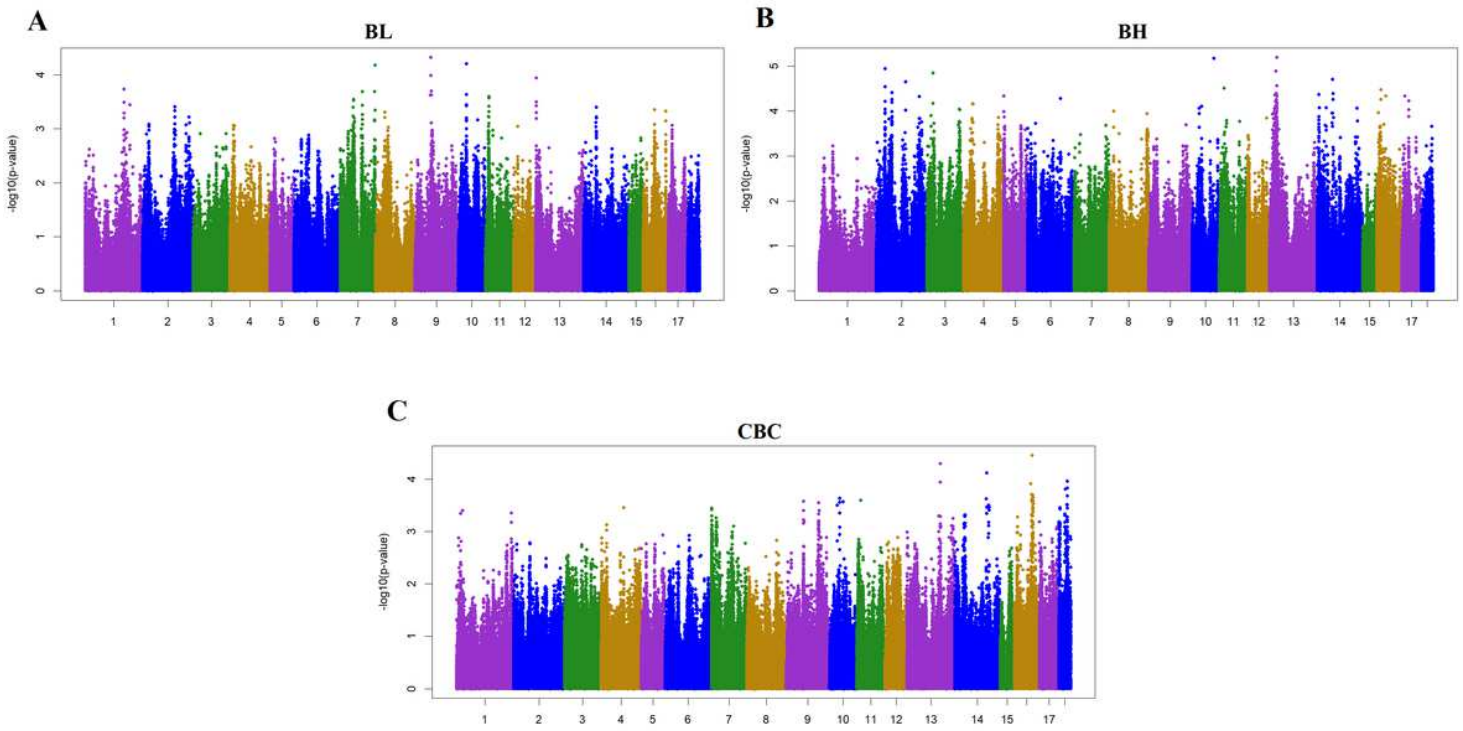

Figure 2

Manhattan plot of the genome-wide association study on four body size traits by using two-trait model ssGWAS. AC, Abdominal circumference; CC, Chest circumference; RW, Rump width; CW, Chest width. AC and CC are a pair of traits, RW and CW are a pair of traits. In the Manhattan plots, negative log10 P-values of the quantified SNPs were plotted against their genomic positions. The x-axis represents the chromosomes, and the y-axis represents the observed -log10(P-value). Different colors indicate various chromosomes. Each trait has a significant threshold of FDR adjusted, for (A) AC, it was 9.74 × 10−6. Similarly, (B) CC was 1.05 × 10−5, (C) RW was 9.60 × 10−6, and (D) CW was 1.01 × 10-5.

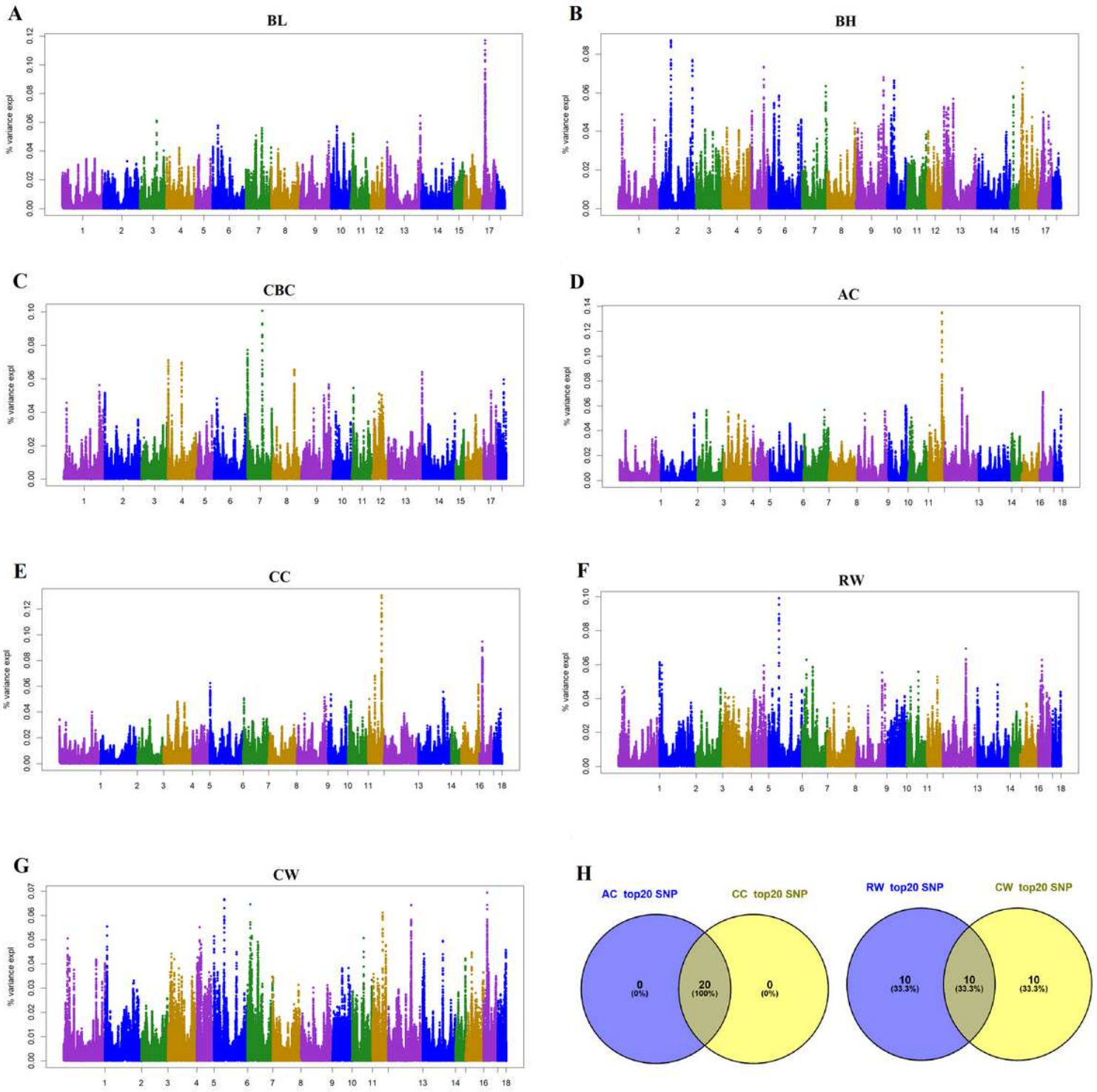

Figure 3

Manhattan plot of the genome-wide association study on seven body size traits and Venn plot of SNPs according to the contribution of SNP to genetic variance by using ssGWAS. BL, Body length; BH, Body height; CBC, Cannon bone circumference; AC, Abdominal circumference; CC, Chest circumference; RW, Rump width; CW, Chest width. BL, BH and CBC were single-trait models；AC, CC, RW and CW were two-trait models. AC and CC are a pair of traits, RW and CW are a pair of traits. In the Manhattan plots(A-G), the

proportion of genetic variance of the quantified SNPs were plotted against their genomic positions. The x-axis represents the chromosomes, and the y-axis represents the percentage of SNP explaining the genetic variance. Different colors indicate different chromosomes. Venn plot(H) of SNPs for the two pairs of body size traits, AC and CC, RW and CW are a pair of traits, respectively.

Supplementary Files

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

- SupplementaryTable.docx