# Genome-wide association analyses of multiple traits in Duroc pigs using low-coverage whole-genome sequencing strategy

Ruifei Yang[1,2]†, Xiaoli Guo[1]†, Di Zhu[1]†, Cheng Bian[1], Yiqiang Zhao[1], Cheng Tan[3], Zhenfang Wu[3,4], Yuzhe Wang[1,2]*, Xiaoxiang Hu[1]*, Ning Li[1]

**Affiliations:**

1 State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing, China

2 College of Animal Science and Technology, China Agricultural University, Beijing, China

3 Guangdong Wens Foodstuffs Group, Xinxing, China

4 National Engineering Research Center For Breeding Swine Industry, College of Animal Science, South China Agricultural University, Guangzhou, China

† Ruifei Yang, Xiaoli Guo and Di Zhu contributed equally to this work.

**Corresponding author:**

Xiaoxiang Hu

State Key Laboratory for Agro-biotechnology, China Agricultural University

Beijing 100193, China

Phone +86-10-62733394

Email: huxx@cau.edu.cn

Yuzhe Wang

State Key Laboratory for Agro-biotechnology, China Agricultural University

Beijing 100193, China

Phone +86-10-62731391

23 Email: yuzhe891@cau.edu.cn

24

**Abstract**

High-density markers discovered in large size samples are essential for mapping complex traits at the gene-level resolution for agricultural livestock and crops. However, the unavailability of large reference panels and array designs for a target population of agricultural species limits the improvement of array-based genotype imputation. Recent studies showed very low coverage sequencing (LCS) of a large number of individuals is a cost-effective approach to discover variations in much greater detail in association studies. Here, we performed cohort-wide whole-genome sequencing at an average depth of 0.73× and identified more than 11.3 M SNPs. We also evaluated the data set and performed genome-wide association analysis (GWAS) in 2885 Duroc boars. We compared two different pipelines and selected a proper method (BaseVar/STITCH) for LCS analyses and determined that sequencing of 1000 individuals with 0.2× depth is enough for identifying SNPs with high accuracy in this population. Of the seven association signals derived from the genome-wide association analysis of the LCS variants, which were associated with four economic traits, we found two QTLs with narrow intervals were possibly responsible for the teat number and back fat thickness traits and identified 7 missense variants in a single sequencing step. This strategy (BaseVar/STITCH) is generally applicable to any populations and any species which have no suitable reference panels. These findings show that the LCS strategy is a proper approach for the construction of new genetic resources to facilitate genome-wide association studies, fine mapping of QTLs, and genomic selection, and implicate that it can be widely used for agricultural animal breeding in the future.

## Background

Genome-wide association studies (GWASs) have generated thousands of genetic variants associated with complex traits in human and agricultural species [1, 2]. The mapping resolution lies on the density of genetic markers which perceive linkage disequilibrium (LD) in sufficiently large populations [3, 4]. Several large-scale whole-genome sequencing projects have been completed, [5] which were designed to identify the underlying mechanisms that drive hereditary diseases in human and genomic selection in the breeding of agricultural species [6-8]. Despite the declining cost of sequencing, it is still difficult to accomplish the desired whole-genome sequencing of every object in a large cohort. In this scenario, imputation-based strategies, which impute low-density panels to higher densities, offer an alternative to systematic genotyping or sequencing [9, 10]. To date, array-based genotype imputation has been widely used in agricultural species [11, 12]. The imputation accuracy of this strategy crucially depends on the reference panel sizes and genetic distances from the target population. However, the unavailability of large reference panels and array designs for target populations in agricultural species limits the improvement of array-based genotype imputation [13, 14]. Inaccurate imputations influence the results of follow-up analyses, such as genome-wide association studies (GWAS) and genomic selection (GS).

In the recently-developed methods, low-coverage sequencing (LCS) of a large cohort has been proposed to be more informative than sequencing fewer individuals at high coverage [15-17]. Sample sizes and haplotype diversity could be more critical than sequencing depths in determining genotype accuracy of most segregating sites and in increasing the power of association studies. Overall, LCS has been proven to have higher power for trait mapping

compared to the array-based genotyping method in human [18]. To date, LCS-based genotype imputation has been employed in many studies using various populations and genotyping algorithms [19-23]. Especially, the STITCH imputation algorithm overcomes the barrier of the lack of good reference panels in non-human species and is applicable even in studies with extremely low sequencing depths [19, 24]. This is a promising approach for agricultural animals without large reference panels in the areas of functional genetic mapping and genomic breeding, but there is no such report yet.

In this study, we describe a cost- and time-efficient low-coverage sequencing method to obtain high-density SNP markers in a large Duroc population [25]. We used the LCS data to demonstrate genotyping and imputation can be inferred with high accuracy in nucleus herds using the BaseVar/STITCH method, allowing further genome-wide association and fine-mapping analyses on multiple traits with high resolution. The LCS strategy provides a powerful way for further exploration of functional genes in agricultural animal breeding.

## Results

### Samples and phenotypes

We choose 2,885 Duroc boars provided by Guangdong Wen's Foodstuff Group (Guangdong, China) as the study subjects, which were the same samples in the study by Tan et al. [25], and all the pigs were managed on a single nucleus farm. We obtained measurements of four phenotypes with different heritabilities, including back fat thickness at 100 kg (BF), loin muscle area at 100 kg (LMA), lean meat percentage at 100 kg (LMP), and teat number (TN). The estimates of genomic narrow-sense heritability were 0.37±0.05, 0.41±0.05, 0.42±0.06 and

89 0.37±0.05, respectively. The phenotypic values followed a near bell-shaped distribution (Figure
90 S1 and Table S1).

91 **Genome sequencing and data acquisition**

92 A Tn5-based protocol was used to prepare sequencing libraries of each pig at low cost (reagent
93 cost: 2.60 $/library) as described in the Materials and Methods. At the beginning, the libraries
94 were sequenced on an Illumina (PE 150) (192 libraries on 2 lanes) or a BGI platform (PE 100)
95 (84 libraries on one lane), and the sequencing depths were 0.40±0.05×/pig for one lane and
96 0.45±0.06×/pig for the other lane on the Illumina platform and 0.66±0.16×/pig on the BGI
97 platform. The results generated by the BGI platform had lower PCR duplicates (2.23%), higher
98 good index reads (97.10%), and higher genome coverage (98.55%) than the Illumina dataset,
99 which had 10.82% of PCR duplicates, 93.64% of good index reads, and 98.50% of genome
100 coverage. The high PCR duplicates would cause a greater number of useless data, leading to a
101 lower depth for each individual pig. Therefore, the remaining samples were all sequenced on
102 the BGI MGISEQ-2000 platform (96 samples/2 lanes). Overall, the total output of the 2869
103 boars approached 5.32 TB, and the majority (96.74%) of reads were successfully mapped to the
104 pig reference genome Sscrofa11.1. Each animal was sequenced at an average of depth of
105 0.73±0.17×, and all the samples had lower levels of PCR duplicates on the BGI platform
106 (2.60±0.08%). Moreover, we also re-sequenced 37 Duroc boars (the core boars of this
107 population) at a high depth (average 10×/per sample), which were used for downstream
108 accuracy evaluation.

109 **Processing pipeline of the low-coverage strategy and accuracy evaluation**

Previous standard methods for joint SNP calling, such as those implemented in GATK and Samtools, were mainly used in high-depth resequencing methods. However, due to the low depth of each base, erroneous SNPs and genotypes could be called using such methods, especially for the GATK HaplotypeCaller algorithm [26]. In this study, we applied the BaseVar algorithm [27] to call SNP variants and estimate allele frequencies, and used STITCH [19] to impute SNPs. The initial screening of chromosome 18 (SSC18) in 1985 samples with BaseVar identified 506,452 and 414,160 bi-allelic candidate SNP sites before and after quality control, respectively. Next, we imputed these SNPs using STITCH, and 322,386 SNPs were retained with a high average call rate (98.89%±0.59%) after quality control. Meanwhile, we also used the GATK UnifiedGenotypeCaller algorithm (different from GATK HaplotypeCaller algorithm) and Beagle to analyze the data and compared the two results (Figure 1). The SNPs detected by BaseVar/STITCH were mostly included (99.32%) in the GATK set, which included 570,919 sites and contained 320,199 SNPs overlapping with the BaseVar/STITCH dataset. To evaluate imputation accuracy, we compared the genotypic concordance (GC) and the allele dosages $R^2$ [28] between the genotypes called in the high coverage whole-genome sequencing analyses of the 37 core boars (high-coverage set) and the imputed genotypes in the low-coverage data (LC set). As a result, a relative high-quality genotype set was acquired with less time consumption when K=10 (Figure S2). We then compared the results generated from the GATK-Beagle and BaseVar-STITCH pipelines in parallel. Figure 2 shows that highly accurate genotypes were obtained using the BaseVar-STITCH pipeline ($R^2$=0.92 and GC=0.97) across all allele frequencies, which excelled far beyond the method using GATK and Beagle ($R^2$=0.48 and

GC=0.71). Therefore, we conclude the BaseVar-STITCH pipeline is a suitable variant discovery and imputation method for the LCS strategy (Figure 1).

Previous studies have demonstrated that sequencing a large number samples at a low depth generally provides a better representation of population genetic variations compared to sequencing a limited number of individuals at a higher depth. Here, using the proper detection and imputation techniques, we obtained SNP sets based on different depths and sample sizes. From Figure 2, we can see that almost all variants had high concordance and $r^2$ values ($R^2$>0.91 and GC>0.96) at all depths when the sample size reached 1000. Even at 0.2× sequencing depth, the SNPs were still detected and imputed with high confidence. Therefore, we conclude that both common and low-frequency SNPs (MAF>0.01) can be obtained with high confidence using information from a larger population in the LC strategy, even when the sequencing depth is around 0.2-0.3×.

**Genetic variations and population structure**

After strict parameter filtering in the pipeline (BaseVar-STITCH, Figure 1), we identified 11,348,460 SNPs in 2885 Duroc pigs with high genotype accuracy ($R^2$=0.92 and GC=0.97), and the density is corresponding to 1 SNP per 200 bp in the pig genome (Table S2). The distribution of variants across the whole genome is mostly uniform, which reflects the high robustness of the LC method. Among all the discovered SNPs, 1,524,015 (accounting for 13.43% of all SNPs) are novel to the pig dbSNP database (data from NCBI:GCA_000003025.6 on Jun, 2017). The majority of identified SNPs were located in intergenic regions (51.98%) and intronic regions (36.85%). The exonic regions contained 1.37% of SNPs, including 0.14% missense SNPs.

153 Interrogating the distribution of the 11.35 million variants in this Duroc population revealed
154 several genetic characteristics. A principal component analysis (PCA) of all the pigs manifests
155 that there was no distinct population stratification among the population (Figure S3). The
156 genome-wide allele frequency spectrum was shown in Figure 3a. The average rate of
157 heterozygosity is low, which is 0.31 of the genomes (Figure 3b), suggesting a strong selection
158 for the pure-bred population. Based on the large population with high-density SNPs, we
159 analyzed the LD decay. The result showed the average pairwise LD $r^2$ decreased slowly along
160 with the increase of distance between markers. The average $r^2$ of the whole genome had been
161 decreased to 0.14 when the distance reached 1 Mb, and slight differences in the average $r^2$
162 existed among 18 chromosomes (Figure 3c and Table S3). Overall, using such high-quality and
163 high-density variants, we could obtain more powerful results from GWAS analyses.

164 **Identification of candidate genes by high-resolution mapping QTLs for TN and BF**

165 We identified a subset of 258,662 SNPs that tagged all other SNPs with MAF >0.1% at LD
166 $r^2$ >0.98 for the first-round GWAS (Table S2). Fine-mapping was performed within 10 Mb of
167 the SNPs to reach 5% Bonferroni-corrected significance threshold genome-wide. Overall, we
168 discovered a total of seven QTLs for the four traits at 5% significance threshold (Figure 4 and
169 Figure S4). The widths of all QTLs' intervals ranged from 40 Kb to 3 Mb; the intervals of five
170 QTLs were more than 2 Mb in width, which was strongly influenced by the local linkage
171 disequilibrium level of this population. In the subsequent analyses, we focused on QTLs
172 containing small numbers of genes (TN and BF's QTLs on SSC7, Figure 4a and 4b).
173 The QTL on SSC7 that has major effect on TN has been widely identified in several commercial
174 breeding lines and hybrids. Our GWAS results show a strong QTL in the same region (Figure

4a). Fine-mapping discovered two narrow LD blocks (SSC7:97.56-97.65 Mb and 98.06-98.10 Mb) containing four candidate genes (Figure 4c). Comparing with the previous results based on the GBS (genotyping-by-sequencing) method, we directly detected 7 missense variants in three genes (*ABCD4*, *PROX2*, and *DLST*). Besides, our result identified the locus of SSC7: 30.24-30.52 Mb was significantly associated with BF. No missense mutation has been detected in this region, but some UTR variants within six genes may have great effects on this trait. All of these genes, *GRM4*, *HMGA1*, *NUDT3*, *RPS10*, *PACSIN1*, and *SPDEF* have been reported to be associated with one or multiple traits in pigs, but clear causal mutations still lack. Our results provide a starting point for further functional investigations.

**Discussion**

To our knowledge, we generated the largest WGS genotyping data set of Duroc population so far, which contains 11 million markers from genotyping 2885 pigs. We expanded the candidate causal mutations for the TN and BF growth-related traits of pigs and demonstrated the efficacy of genetic fine-mapping utilizing low-coverage sequencing in animal populations with unavailable reference panels. This method is expected to have widespread usages in genome-wide association studies, fine mapping of QTLs, and genomic selection.

This study identified an optimal design, taking into account the number of samples, sequencing depth, and imputation algorithm. Two critical data can be referenced for future research on animals without large reference panels: the BaseVar-STITCH pipeline allows the GC higher than 0.96 when the sample size of 1000 and the sequencing depth of 0.2× were reached, or when the sample size of 500 and the sequencing depth of 0.5× were reached. The GC values under both conditions are significantly higher than other studies of array-based genotype

197 imputation. We also found that the genotype accuracy is more sensitive to sample sizes than
198 sequencing depths. In other words, the results demonstrated low-coverage designs are more
199 powerful than deep sequencing of fewer individuals for animal sequencing studies.
200 The QTL region on SSC7, which was identified in the current study, has also been reported to
201 be associated with TN, the number of vertebrae (NVE), or the number of ribs by GWASs [25,
202 29-31]. Vertebrae develop from the somites, whose ventral elongation also determines the
203 correct dorsoventral position of mammary epithelium along the flank [32]. Thus, somites may
204 be the progenitor cells of vertebrae, ribs, and mammary glands, and the variations in the genes
205 downstream of the developmental cascade for the formation of the mammary gland are most
206 likely responsible for the QTL we detected for TN in the GWAS [33]. It is worth noting that
207 five missense variants are discovered in *PROX2*, which is one of the vertebrate homologs of
208 Drosophila melanogaster homeodomain-containing protein Prospero, and may be involved in
209 the determination of cell fate and the establishment of the body plan [34]. We suggest these
210 missense mutations may be the causal variants for the phenotype, although functional studies
211 are needed to validate this hypothesis. For the QTL associated with BF, we found three UTR
212 variants located in *HMGA1*. *HMGA1* is a promising candidate gene associated with growth,
213 carcass, organ weights, fat metabolism, as it has been reported to involve in a variety of genetic
214 pathways regulating cell growth and differentiation, glucose uptake, and white and brown
215 adipogenesis [35-39]. Overall, the QTLs with narrow intervals and a few candidate genes were
216 identified, which emphasizes the potential of identifying new mutations in QTLs using the low-
217 coverage sequencing method in a single sequencing step.

Increasing the marker density was proposed to have the potential to improve the accuracy of genomic prediction for quantitative traits [40]. However, in recent studies, SNP chips were mostly used to build genetic relationship matrices [41, 42], which could not catch all recombination events in a given population. Here, the whole-genome low-coverage sequencing data gave the best accuracy of prediction, since most causal mutations that underlie a trait are expected to be included. Meanwhile, the haplotype reference panel can accommodate new haplotypes due to recombination at any time, thus improving the issue of the decrease of prediction accuracy over generations. Our data can cover the sites of various of SNP chips well because the genome coverage exceeds 98.36%, and it is competitive with arrays in terms of cost and SNP density. Besides, most researchers or breeders may concern more about the efficiency of the method. The development of application servers brings hope to solve the time-consuming computational issue of genotyping using the whole genome sequencing data. In this study, we applied GTX, which is an FPGA-based hardware accelerator platform [43], to do the alignments, and all 3000 alignments were accomplished in two days. Then the genotyping and imputation could be achieved on the cluster server or even cloud server in a single day. Therefore, the accuracy and timeliness issue for genomic prediction could be all resolved in the near future. An alternative solution at present is that we can select different useful tag-SNPs to make ultra-low-density SNP chips for various traits with different genetic architectures using the high-density genetic map built by LC data, since all possible haplotypes were available in the haplotype database. Further, the cost could be reduced, and breeding could be achieved more efficiently.

## Methods

240 **Ethics Statement**

241 All procedures involving animals in this study were carried out in accordance with the
242 guidelines for the care and use of experimental animals established by the Ministry of Science
243 and Technology of the People's Republic of China (Approval Number: 2006-398). All the
244 animal experiment protocols were approved by the Animal Welfare Committee of China
245 Agricultural University (Permission Number: SKLAB-2014-04-02).

246 **Animals, phenotyping, and DNA Extraction**

247 The Duroc boars used for this study were provided by Guangdong Wen's Foodstuff Group
248 (Guangdong, China), which were born from September 2011 to September 2013. All pigs were
249 managed on a single nucleus farm. The associated phenotype data used in this study included
250 teat number (TN), back fat thickness at 100 kg (BF), loin muscle area at 100 kg (LMA), and
251 lean meat percentage at 100 kg (LMP). The last three phenotypes were recorded when the
252 weights of pigs reached 100 ± 5 Kg. The phenotype data of TN were acquired from Tan's study,
253 and BF, LMA, and ELMP were measured over the last three to four ribs using a b-ultrasound-
254 scan equipment (Aloka SSD-500). The phenotypic values of TN followed a near bell-shaped
255 distribution, which is same as reported by Tan et al., and the data of other three phenotypes all
256 nearly followed the normal distribution (Fig S1). In addition, body weights were recorded at
257 birth, and at the beginning (30 ± 5 Kg) and the end (100 ± 5 Kg) of the experiment. Genomic
258 DNA was extracted from the ear tissue using a DNeasy Blood & Tissue Kit (Qiagen 69506),
259 assessed using a NanoDrop, and checked in a 1% agarose gel. All the samples were quantified
260 using a Qubit 2.0 Fluorometer, and then diluted to 40 ng/ml in 96-well plates.

261 **Tn5 Libraries generation and sequencing**

Equal amounts of Tn5ME-A/Tn5MErev and Tn5ME-B/Tn5MErev were incubated at 72 ℃ for 2 min, then were placed on ice immediately. Tn5 (Karolinska Institutet 171 77 Stockholm, Sweden) was loaded with the Tn5ME-A+rev and Tn5ME-B+rev in 2× Tn5 dialysis buffer at 25℃ for 2 h. All linker oligonucleotides were same as the previous report [44].

Tagmentation were carried out at 55℃ for 10 minutes by mixing 4 μl 5×TAPS-$MgCl_2$, 2 μl dimethylformamide (DMF) (Sigma Aldrich), 1 μl of the Tn5 that pre-diluted to 16.5 ng/μl, 50 ng DNA, and nuclease-free water. The total volume of the reaction was 20 μl. Then 3.5 μl 0.2% SDS was added, and Tn5 was inactivated for another 10 min at 55℃.

KAPA HiFi HotStart ReadyMix (Roche) was used for PCR amplification. The primers were designed for MGI sequencers, with the reverse primers contained 96 different index adaptors to distinguish individual library. The PCR program was as follows: 9 min at 72℃, 30 sec at 98℃, and then 9 cycles of 30 sec at 98℃, 30 sec at 63℃, followed by 3 min at 72℃. The products were quantified by Qubit Fluorometric Quantitation (Invitrogen), then the groups of 96 indexed samples were pooled with equal amounts.

Size-selection was performed using the AMPure XP beads (Beckmann), with the left side size selection ratio was 0.55×, and the right was 0.1×. The final libraries were sequenced on 2 lanes of MGISEQ-2000 to generate 2×100bp paired-end reads or on 1 lane of BGISEQ-500 to generate 2×100 bp paired-end reads.

**High depth sequencing of 37 boars**

We sequenced 37 out of the total 1985 pigs using the Hiseq X Ten system at a high depth of 10×. GTX by Genetalks company, a commercially available FPGA-based hardware accelerator platform, was used in this study for both mapping clean reads to the Sscrofa11.1 reference

284 genome (ftp://ftp.ensembl.org/pub/release-97/fasta/sus_scrofa/dna/) and variant calling. The
285 alignment process is accelerated by FPGA implementation of a parallel seed-and-extend
286 approach based on the Smith-Waterman algorithm, while the variant calling process is
287 accelerated by FPGA implementation of GATK HaplotypeCaller (PairHMM). GATK multi-
288 sample best practice [45] was used to call and genotype SNPs for the 37 pigs, and the SNPs
289 were hard filtered with a relatively strict option "QD < 10.0 || ReadPosRankSum < -8.0 || FS >
290 10.0 || MQ<40.0". The average running time from a fastq file to a gvcf file was about 3 min for
291 each sample in this study.

292 **Low coverage sequencing data analyses**

293 Sequencing reads from the low coverage samples were mapped to Sscrofa11.1 reference
294 genome using GTX-align, which includes a step of marking PCR duplicates. The indel
295 realignment and base quality recalibration modules in GATK were applied to realign the reads
296 around indel candidate loci and to recalibrate the base quality. Variant calling was done using
297 the BaseVar and hard filtered with EAF >= 0.01 and the Depth that is greater than or equal to
298 1.5 times InterQuartile Range. The detailed BaseVar algorithm to call SNP variants and to
299 estimate allele frequency was described in a pevious report [27]. We used STITCH [19] to
300 impute genotype probabilities for all individuals. The key parameter K (number of ancestral
301 haplotypes) was decided based on the tests in SSC18. Results were filtered with an imputation
302 info score > 0.4 and Hardy-Weinberg Equilibrium (HWE) p-value > 1e−6. Two validation
303 actions were taken to calculate the accuracy of imputation. The first parameter is genotypic
304 concordance (GC), which was calculated as the number of correctly-imputed genotypes divided

by total sites. Another parameter is allele dosages $R^2$, which was described in a previous report [28]. The SNPEff program [46] was used to annotate the variants.

**Heritability estimation and Genome-wide association**

Heritability was estimated using a mixed model as the following:

$$\mathbf{y} = \mathbf{X_b b} + \mathbf{Za} + \mathbf{Zc} + \mathbf{e}$$

with $Var(\mathbf{y}) = \mathbf{ZA_aZ'}\sigma_a^2 + \mathbf{I}\sigma_c^2 + \mathbf{I}\sigma_e^2$, where Z is an incidence matrix allocating phenotypic observations to each animal; **b** is the vector of fixed year-month effects for BF, LMA, and ELMP; **b** also includes birth weight, the weights at the beginning and end of the test as covariance; $\mathbf{X_b}$ is the incidence matrix for **b**; **a** is the vector of additive values based on the pedigree information; **c** is the vector of random family effects; $\mathbf{A_a}$ is a pedigree-based additive relationship matrix; $\sigma_a^2$ is the additive variance; $\sigma_c^2$ is the variance of random family effects; and $\sigma_e^2$ is the residual variance. Variance components for BF, LMA, and LMP were estimated by AIREMlF90 program, and by thrgibbs1f90 program for TN. Both programs were in the BLUPF90 package. The additive heritability was defined as: $h_a^2 = \sigma_a^2 /(\sigma_a^2 + \sigma_c^2 + \sigma_e^2)$.

A subset of 258,662 SNPs that tagged all other SNPs with MAF >1% at LD $r^2$ <0.98 and the call rate >95% were retained for genome-wide association analysis. A mixed linear model (MLM) approach was used for the genome-wide association analyses as implemented in the GCTA package (v1.24) [47]. The statistical model during analyses of TN included the year and season as discrete covariates. For BF, LMA, and ELMP, the year and season were included as discrete covariates, and birth weight, the weight at the beginning and end of the test were used as quantitative covariate. To correct multiple testing across the genome, a Bonferroni correction

was applied to compensate for the number of estimated independent markers from a PCA analysis and was performed as follows. A subset of SNPs that were in approximate linkage equilibrium with each other was obtained by removing one in each pair of SNPs if the LD was greater than 0.5 using the PLINK v1.07 '--indep-pairwise' command [48]. The squared correlation coefficient ($r^2$) between the genotypes was calculated using the vcftools '--geno-r2' command [45]. Consequently, for our population, the genome-wide 1% significance threshold was determined as p-value < $3.47 \times 10^{-7}$ (0.01/28,828), and a suggestive association was determined as $1.73 \times 10^{-6}$ (0.05/28,828).

## Declarations

## Competing interests

The authors declare that they have no competing interests.

## Funding

This project was supported by the 948 Program of the Ministry of Agriculture of China (2012-G1(4)), the National Transgenic Grand Project [2016ZX08009003-006(2016-2018), (2016ZX08006002-003)], and the Science and Technology Innovation Strategy Projects of Guangdong Province [2019B020203002].

## Acknowledgements

We thank Siyang Liu and Xun Xu for their valuable suggestions on data analyses. Zhuo Song, Chungen Yi, and Wenjuan Wei provided the services of FPGA-based hardware accelerator

platform and the Batch Compute system in Aliyun cloud. We also thank Zhaoliang Liu for improving the manuscript. Part of the analysis was performed on the high-performance computing platform of the State Key Laboratory of Agrobiotechnology.

## References

1. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7-24.

2. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet. 2010;42(11):961-7.

3. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11(7):499-511.

4. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007;39(7):906-13.

5. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.

6. Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 2010;42(12):1053-9.

7. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46(8):858-65.

8. Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. Annu Rev Anim Biosci. 2019;7:89-102.

9. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the

367 next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529.

368 10. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype

369 imputation in genome-wide association studies through pre-phasing. Nature Genetics. 2012;44(8):955-

370 +.

371 11. Yan G, Qiao R, Zhang F, Xin W, Xiao S, Huang T, et al. Imputation-Based Whole-Genome

372 Sequence Association Study Rediscovered the Missing QTL for Lumbar Number in Sutai Pigs. Sci Rep.

373 2017;7(1):615.

374 12. van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I, et al. Accuracy of

375 imputation to whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2014;46:41.

376 13. van den Berg S, Vandenplas J, van Eeuwijk FA, Bouwman AC, Lopes MS, Veerkamp RF.

377 Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide

378 association studies. Genetics Selection Evolution. 2019;51.

379 14. Swarts K, Li HH, Navarro JAR, An D, Romay MC, Hearne S, et al. Novel Methods to Optimize

380 Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. Plant

381 Genome-Us. 2014;7(3).

382 15. Buerkle CA, Gompert Z. Population genomics based on low coverage sequencing: how low should

383 we go? Mol Ecol. 2013;22(11):3028-35.

384 16. Huang L, Wang B, Chen RT, Bercovici S, Batzoglou S. Reveel: large-scale population genotyping

385 using low-coverage sequencing data. Bioinformatics. 2016;32(11):1686-96.

386 17. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for

387 design of complex trait association studies. Genome Res. 2011;21(6):940-51.

388 18. Gilly A, Southam L, Suveges D, Kuchenbaecker K, Moore R, Melloni GEM, et al. Very low-depth

389 whole-genome sequencing in complex trait association studies. Bioinformatics. 2019;35(15):2555-61.

390 19. Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence without reference

391 panels. Nature Genetics. 2016;48(8):965-+.

392 20. Ros-Freixedes R, Gonen S, Gorjanc G, Hickey JM. A method for allocating low-coverage

393 sequencing resources by targeting haplotypes rather than individuals. Genetics Selection Evolution.

394 2017;49.

395 21. Fragoso CA, Heffelfinger C, Zhao HY, Dellaporta SL. Imputing Genotypes in Biallelic Populations

396 from Low-Coverage Sequence Data. Genetics. 2016;202(2):487-+.

397 22. Bickhart DM, Hutchison JL, Null DJ, VanRaden PM, Cole JB. Reducing animal sequencing

398 redundancy by preferentially selecting animals with low-frequency haplotypes. J Dairy Sci.

399 2016;99(7):5526-34.

400 23. Zan Y, Payen T, Lillie M, Honaker CF, Siegel PB, Carlborg O. Genotyping by low-coverage whole-

401 genome sequencing in intercross pedigrees from outbred founders: a cost-efficient approach. Genet Sel

402 Evol. 2019;51(1):44.

403 24. Nicod J, Davies RW, Cai N, Hassett C, Goodstadt L, Cosgrove C, et al. Genome-wide association

404 of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nat Genet.

405 2016;48(8):912-8.

406 25. Tan C, Wu ZF, Ren JL, Huang ZL, Liu DW, He XY, et al. Genome-wide association study and

407 accuracy of genomic prediction for teat number in Duroc pigs using genotyping-by-sequencing. Genetics

408 Selection Evolution. 2017;49.

409 26. Ros-Freixedes R, Battagin M, Johnsson M, Gorjanc G, Mileham AJ, Rounsley SD, et al. Impact of

410 index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage

411 sequencing. Genet Sel Evol. 2018;50(1):64.

412 27. Liu S, Huang S, Chen F, Zhao L, Yuan Y, Francis SS, et al. Genomic Analyses from Non-invasive
413 Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population
414 History. Cell. 2018;175(2):347-59 e14.

415 28. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase
416 inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84(2):210-23.

417 29. Duijvesteijn N, Veltmaat JM, Knol EF, Harlizius B. High-resolution association mapping of number
418 of teats in pigs reveals regions controlling vertebral development. BMC Genomics. 2014;15:542.

419 30. Verardo LL, Silva FF, Lopes MS, Madsen O, Bastiaansen JW, Knol EF, et al. Revealing new
420 candidate genes for reproductive traits in pigs: combining Bayesian GWAS and functional pathways.
421 Genet Sel Evol. 2016;48:9.

422 31. Ren DR, Ren J, Ruan GF, Guo YM, Wu LH, Yang GC, et al. Mapping and fine mapping of
423 quantitative trait loci for the number of vertebrae in a White Duroc x Chinese Erhualian intercross
424 resource population. Anim Genet. 2012;43(5):545-51.

425 32. Veltmaat JM, Relaix F, Le LT, Kratochwil K, Sala FG, van Veelen W, et al. Gli3-mediated somitic
426 Fgf10 expression gradients are required for the induction and patterning of mammary epithelium along
427 the embryonic axes. Development. 2006;133(12):2325-35.

428 33. van Son M, Lopes MS, Martell HJ, Derks MFL, Gangsei LE, Kongsro J, et al. A QTL for Number
429 of Teats Shows Breed Specific Effects on Number of Vertebrae in Pigs: Bridging the Gap Between
430 Molecular and Quantitative Genetics. Front Genet. 2019;10.

431 34. Pistocchi A, Bartesaghi S, Cotelli F, Del Giacco L. Identification and expression pattern of zebrafish
432 prox2 during embryonic development. Dev Dyn. 2008;237(12):3916-20.

433 35. Gong H, Xiao S, Li W, Huang T, Huang X, Yan G, et al. Unravelling the genetic loci for growth and
434 carcass traits in Chinese Bamaxiang pigs based on a 1.4 million SNP array. J Anim Breed Genet.
435 2019;136(1):3-14.
436 36. Liu X, Wang LG, Liang J, Yan H, Zhao KB, Li N, et al. Genome-Wide Association Study for Certain
437 Carcass Traits and Organ Weights in a Large WhitexMinzhu Intercross Porcine Population. J Integr Agr.
438 2014;13(12):2721-30.
439 37. Arce-Cerezo A, Garcia M, Rodriguez-Nuevo A, Crosa-Bonell M, Enguix N, Pero A, et al. HMGA1
440 overexpression in adipose tissue impairs adipogenesis and prevents diet-induced obesity and insulin
441 resistance. Sci Rep. 2015;5:14487.
442 38. Wang LG, Zhang LC, Yan H, Liu X, Li N, Liang J, et al. Genome-Wide Association Studies Identify
443 the Loci for 5 Exterior Traits in a Large White x Minzhu Pig Population. Plos One. 2014;9(8).
444 39. Ji JX, Yan GR, Chen D, Xiao SJ, Gao J, Zhang ZY. An association study using imputed whole-
445 genome sequence data identifies novel significant loci for growth-related traits in a Duroc x Erhualian
446 F-2 population. Journal of Animal Breeding and Genetics. 2019;136(3):217-28.
447 40. Meuwissen T, Goddard M. Accurate prediction of genetic values for complex traits by whole-
448 genome resequencing. Genetics. 2010;185(2):623-31.
449 41. Song H, Zhang J, Jiang Y, Gao H, Tang S, Mi S, et al. Genomic prediction for growth and
450 reproduction traits in pig using an admixed reference population. J Anim Sci. 2017;95(8):3415-24.
451 42. Zhang C, Kemp RA, Stothard P, Wang Z, Boddicker N, Krivushin K, et al. Genomic evaluation of
452 feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants.
453 Genet Sel Evol. 2018;50(1):14.
454 43. Xing Y, Li G, Wang Z, Feng B, Song Z, Wu C. GTZ: a fast compression and cloud transmission

455 tool optimized for FASTQ files. BMC Bioinformatics. 2017;18(Suppl 16):549.

456 44. Picelli S, Bjorklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and

457 tagmentation procedures for massively scaled sequencing projects. Genome Res. 2014;24(12):2033-40.

458 45. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format

459 and VCFtools. Bioinformatics. 2011;27(15):2156-8.

460 46. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and

461 predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila

462 melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80-92.

463 47. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis.

464 Am J Hum Genet. 2011;88(1):76-82.

465 48. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for

466 whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75.

467

468

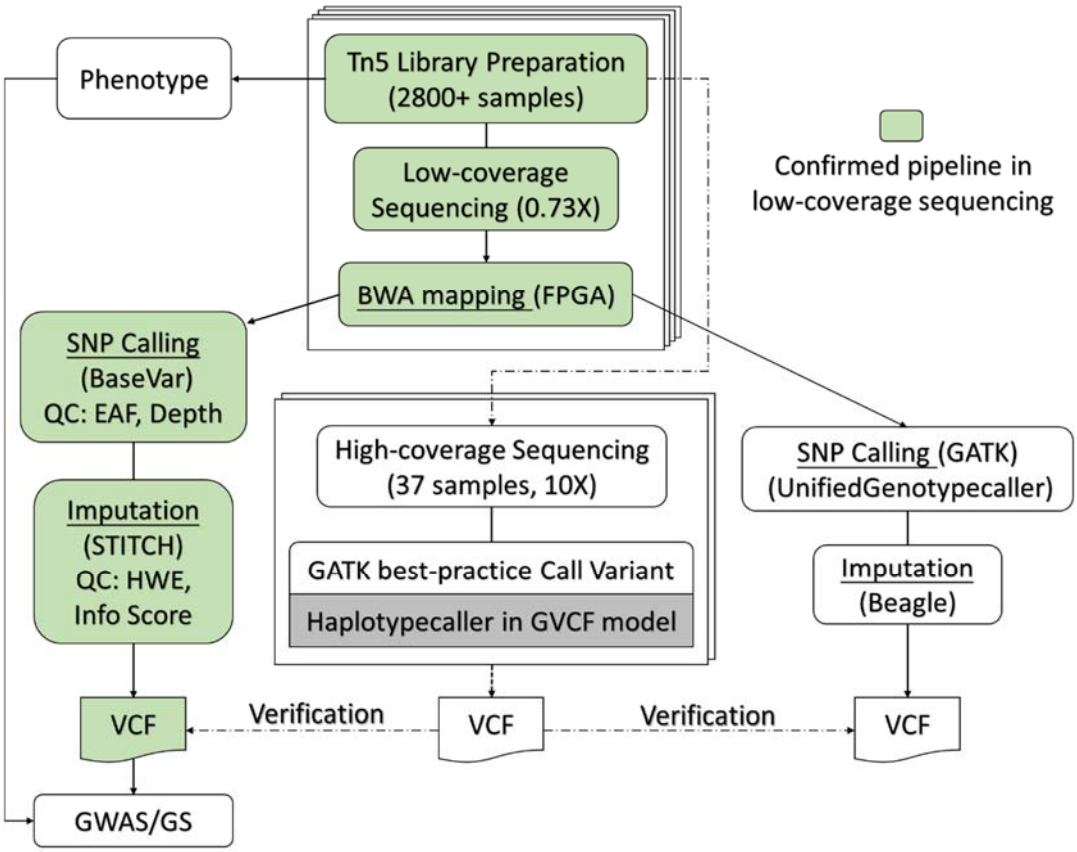



**Figure 1. The LCS study design.** The flow chart summarizes the steps used to detect and impute SNPs, where the green block represents the pipeline for the LCS analysis (BaseVar-STITCH). The data generated from the GATK-Beagle pipeline were compared with that of the BaseVar-STITCH pipeline, and the data generated from the high-coverage sequencing analyses were used to verify the above results. The BaseVar-STITCH pipeline was used in the further GWAS study.

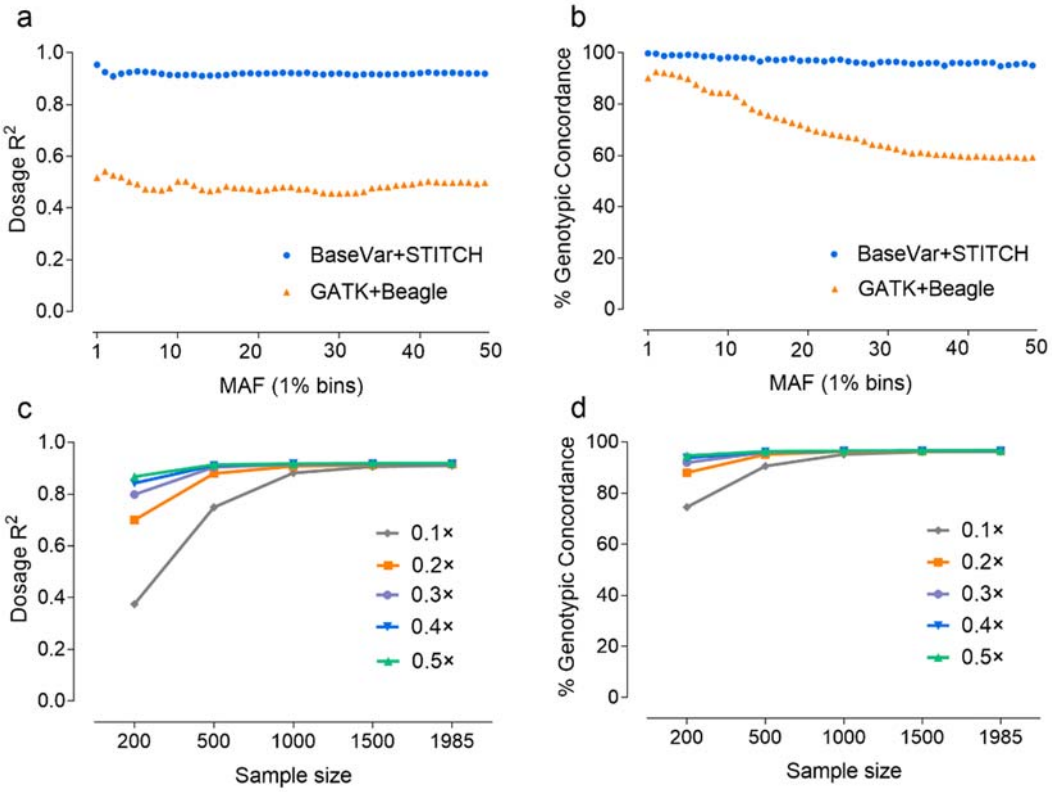
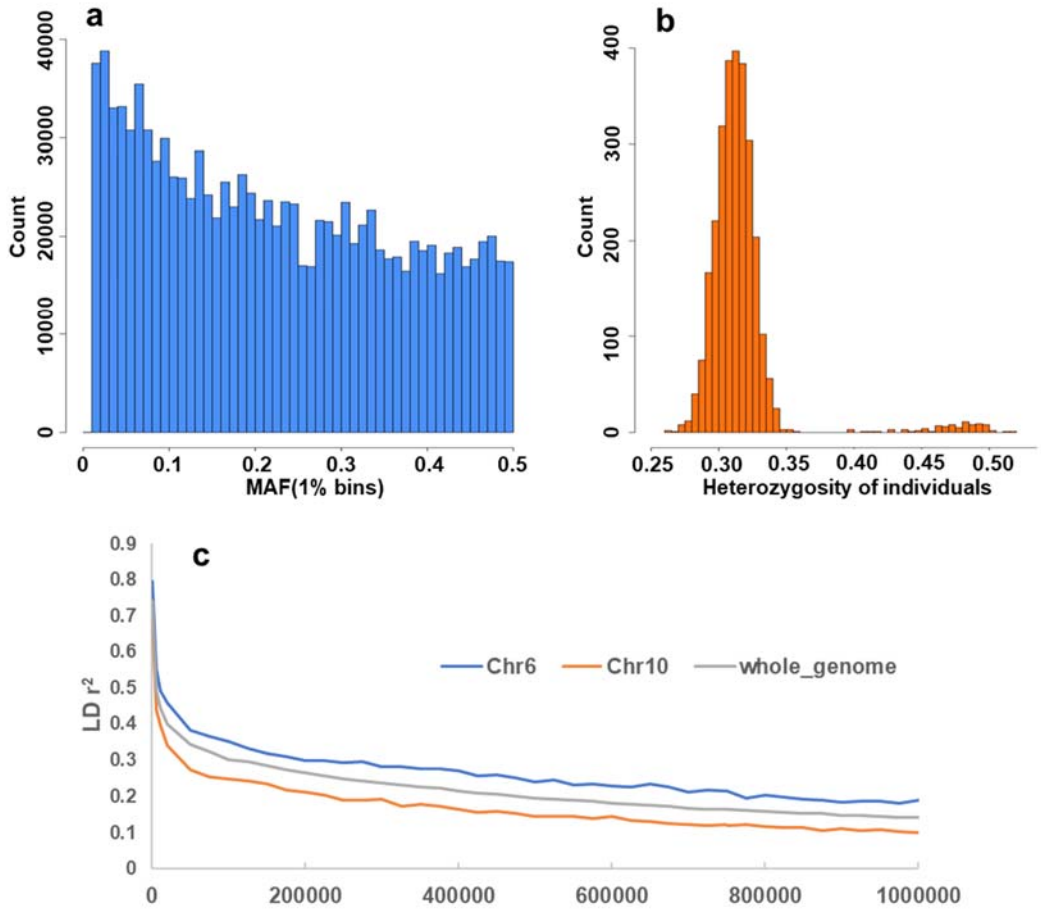



**Figure 2. Dosage $R^2$ and genotypic concordance (%) values for different MAFs and sample sizes.** (a) and (b) show the comparison of Dosage $R^2$ and genotypic concordance values between the BaseVar/STITCH for LGS (blue) and the GTAK/Beagle (orange) pipelines, and (c) and (d) show the comparison of Dosage $R^2$ and genotypic concordance values among different sequencing depths.




484

485 **Figure 3. Sequencing diversity of the Duroc population.** (a) Histogram of allele counts by

486 each 1% MAF bin. (b) Histogram of genome-wide heterozygosity. (c) The extent of linkage

487 disequilibrium (LD), in which the LD on chromosome 6 and 10 represent the highest and lowest

488 levels among the whole genome respectively.

489

490

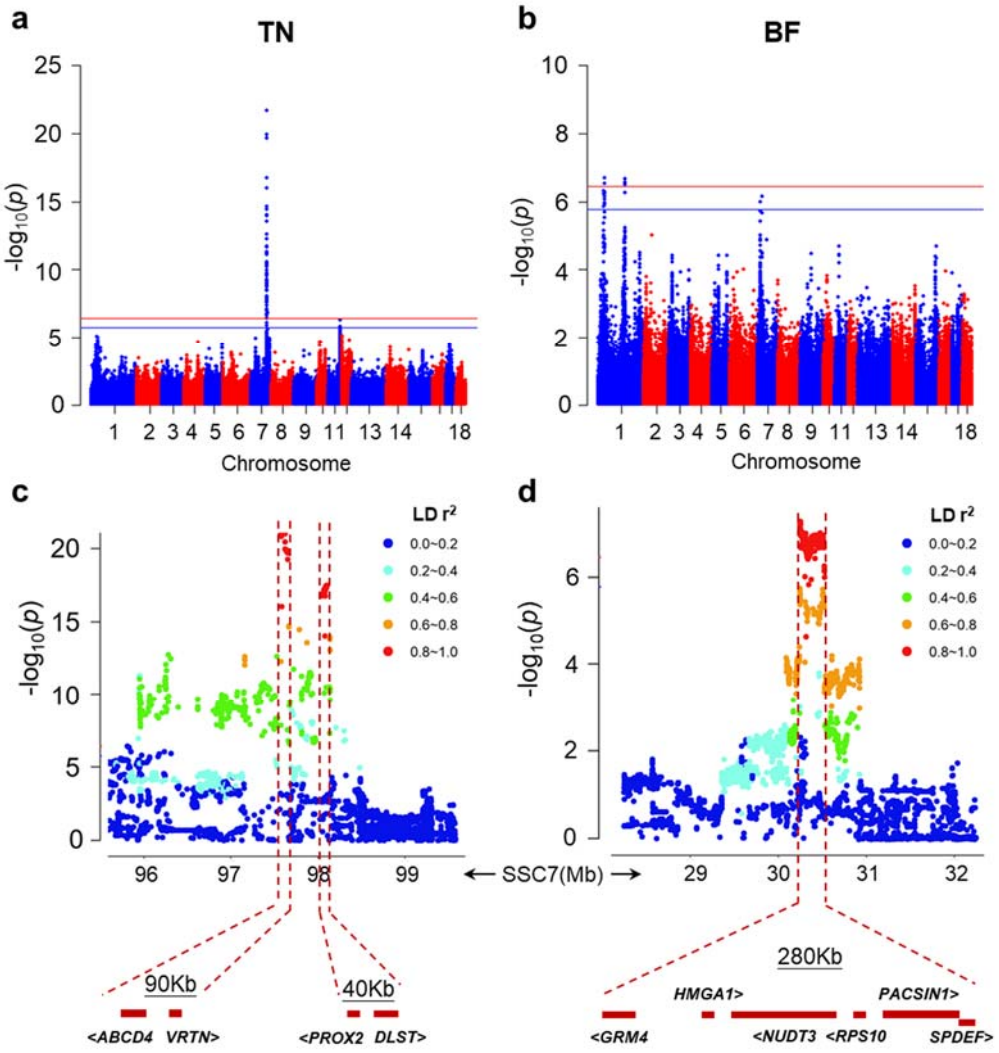



**Figure 4. Manhattan plots and fine-mapping of total tit number (TN) and back fat thickness (BF).** (a) and (b) depict the TN and BF association signals on the whole genome, in which the blue and red horizontal line represent the 0.05 ($p < 1.73\times10^{-6}$) and 0.01 ($p < 3.46\times10^{-7}$) significant levels after Bonferroni correction. (c) Fine-mapping of TN using the entire set of SNPs, in which two isolated regions on chromosome 7 with the lengths of 90 Kb and 40 Kb were detected as QTLs. (d) Fine-mapping of BF using the entire set of SNPs, a narrow QTL with the length of 280 Kb on chromosome 7 was detected.