Optimizing Genomic Selection in Dezhou Donkey Using Low Coverage Whole Genome Sequencing

Changheng Zhao
Shandong Agricultural University
Jun Teng
Shandong Agricultural University
Xinhao Zhang
Shandong Agricultural University
Dan Wang
Shandong Agricultural University
Xinyi Zhang
Shandong Agricultural University
Shiyin Li
Shandong Agricultural University
Haijing Li
Dong-E E-Jiao Co.
Xin Jiang
Shandong Agricultural University
Chao Ning
Shandong Agricultural University
Qin Zhang (qzhang@cau.edu.cn)
Shandong Agricultural University

Research

Keywords: Dezhou donkey, Low coverage whole genome sequencing, Genotype imputation, Genomic selection

Posted Date: June 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-607740/v1

Optimizing Genomic Selection in Dezhou Donkey Using low coverage whole genome sequencing

Changheng Zhao[1], Jun Teng[1], Xinhao Zhang[1,2], Dan Wang[1], Xinyi Zhang[1], Shiyin Li[1], Xin Jiang[1], Haijing Li[2], Chao Ning[1*], Qin Zhang[1*]

1. Shandong Provincial Key Laboratory of Animal Biotechnology and Disease Control and Prevention, College of Animal Science and Veterinary Medicine, Shandong Agricultural University, Tai'an, 271018, China

2. National Engineering Research Center for Gelatin-based TCM, Dong-E E-Jiao Co., Ltd, 78 E-Jiao Street, Donge County, 252201, Shandong Province, China

[*] Corresponding author:

Chao Ning

ningchao@sdau.edu.cn

Qin Zhang

qzhang@sdau.edu.cn

Abstract

Background: Low coverage whole genome sequencing is a low-cost genotyping technology. Combining with genotype imputation approaches, it is likely to become a critical component of cost-efficient genomic selection programs in agricultural livestock. Here, we used the low-coverage sequence data of 617 Dezhou donkeys to investigate the performance of genotype imputation for low coverage whole genome sequence data and genomic selection based on the imputed genotype data. The specific aims were: (i) to measure the accuracy of genotype imputation under different sequencing depths, sample sizes, MAFs, and imputation pipelines; and (ii) to assess the accuracy of genomic selection under different marker densities derived from the imputed sequence data, different strategies for constructing the genomic relationship matrixes, and single- vs multi-trait models.

Results: We found that a high imputation accuracy (> 0.95) can be achieved for sequence data with sequencing depth as low as 1x and the number of sequenced individuals equal to 400. For genomic selection, the best performance was obtained by using a marker density of 410K and a **G** matrix constructed using marker dosage information. Multi-trait GBLUP performed better than single-trait GBLUP.

Conclusions: Our study demonstrates that low coverage whole genome sequencing would be a cost-effective method for genomic selection in Dezhou Donkey.

Keywords: Dezhou donkey, Low coverage whole genome sequencing, Genotype imputation, Genomic selection

Background

Dezhou Donkey, originated from Dezhou area, Shandong Province, China, is one of major donkey breeds in China. It is famous for its large body size (thus good meat production ability) and excellent skin quality (for producing donkey-hide gelatin). It has been introduced as breeding stock into many provinces and cities, and has also brought considerable economic benefits to farmers. Therefore, Dezhou Donkey plays an important role in the donkey industry in China. However, selective breeding in the sense of modern animal breeding theory has long been ignored for any donkey breeds in China. In recent years, along with the increasing of the importance of the donkey industry in livestock agriculture in China, donkey breeding is gradually becoming an important issue in donkey production and some breeding work are carrying out in Dezhou Donkey population.

Starting with the pioneered work of Meuwissen et al. [1], genomic selection (GS) has been widely used in selective breeding in almost all major farm animal species, and has brought great increasement of genetic progresses and economic benefit for many animal breeding industries [2-4] . Typically, GS is carried out using a high (or medium) density marker (SNP) array. Many commercial SNP arrays have been developed for almost all major farm animal species. However, there is no such array for donkey, which inhibits the application of GS in donkey.

Recently, along with the rapid development of next generation sequencing technology and reduction of sequencing cost, GS using genotypes revealed by whole genome sequencing (WGS, instead of SNP array) has drawn interests of animal GS

community with the motivation of further improving the selection accuracy, better application of GS across breeds/populations, and better persistence of accuracy across generations [5-6]. To capture all variants in the genome, a sequencing depth of about 10x is generally required. However, at present 10x sequencing is still too expensive for large scale GS application. An alternative is to perform low coverage whole genome sequencing (lcWGS) at only about 1x, and then recovering the missing genotypes by imputation to ensure that all individuals have genotypes for a shared set of variants. This approach has been used in human and some animal species for genome-wide association study and genomic selection/prediction and approved to be a feasible alternative to normal sequencing [7-10]. Since the cost of lcWGS can even be lower than that of SNP array, it is considered as a cost-effective genotyping approach for GS and GS based this approach was referred as GS 2.0 by Hickey (2013) [11].

A critical issue of lcWGS-based GS is the accuracy of imputation of missing genotypes, which is affected by several factors, such as sequencing depth, population size, minor allele frequency (MAF), and imputation method. A number of imputation methods for lcWGS data have been proposed [12-14]. However, most of these methods require a high-density haplotype reference panel, which are not available for most animal species. Davies et al. (2016)[12] proposed a method called STITCH for imputation based only on sequencing read data, without requiring a haplotype reference panel, which provides an opportunity of using lcWGS technology for species that lack a haplotype reference panel.

In this study, we evaluated the imputation accuracy of lcWGS data with respect to

different sequencing depths, population sizes, MAFs, and imputation pipelines using 617 Dezhou Donkey animals which were sequenced with an average depth of 3.5x. We then used the imputed genotypes to investigate the performance of genomic selection for birth weight and weaning weight in the Dezhou Donkey population under different marker densities, strategies for constructing the genomic relationship matrices, and single- vs two-trait models.

Materials and Methods

Animals

Blood samples from 617 Dezhou Donkey animals were collected from a donkey farm in Shandong Province. Total DNA was isolated using the QIAamp DNA Investigator kit (QIAGEN, Hilden, Germany) and following the manufacturer's instruction. DNA quality was evaluated by spectrophotometry and agarose gel electrophoresis.

All experimental chickens were maintained, and all the studies were carried out according to the guideline of the experimental animal management of Shandong Agricultural University (SDAUA-2018-018).

Low coverage whole genome sequencing

DNA templates were ultrasonically sheared using a Covaris E220 (Covaris, Woburn, MA, USA) to yield ~150 bp fragments, and then prepared for sequencing libraries following the workflow of the NEBNext Ultra DNA Library Preparation Protocol. Multiple Ampure Bead XP cleanups (Beckman Coulter, Brea, CA, USA) were

conducted to remove any adapter dimer that might have developed. The quality and concentration of libraries were determined on an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA). The quality-controlled genomic library for each sample was PE150 sequenced using the Illumina NovaSeq 6000 sequencing system. The average sequencing coverage of the sample data was 3.5x.

Read quality was assessed using the FastQC software (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) with focus on base quality scores and GC content, N content, and sequence duplication levels. All data reached a nucleotide length of longer than 50 bp and a Phred quality score of lower than 30, which were aligned to the donkey reference genome [15] by BWA [16]. Samtools [17] was used to transfer formats, sort and index files.

Pipelines for genotype imputation

We compared two imputation pipelines, i.e., Bcftools + Beagle and BaseVar + STITCH. In the first pipeline, we called SNPs using Bcftools[18], performed quality control using PLINK [19] with the parameters of geno (>90%) and MAF (>1%), and then conducted genotype imputation using Beagle v4.1 [20]. In the second pipeline, we called SNPs using BaseVar [7] , filtered with EAF >= 0.01, and then imputed the missing genotypes (with probabilities) using STITCH. The resulted SNP data were filtered with an imputation info_score > 0.4 and a Hardy-Weinberg Equilibrium (HWE) *p*-value > 1e-6.

Evaluation of imputation accuracy

We evaluated the imputation accuracy using the sequencing data of 18 Dezhou Donkey animals, which were sequenced with an average sequencing coverage of 13.5x. Chromosomes 1, 19 and 30 were chosen to compare the imputed and typed genotypes in terms of genotype concordance measured as proportion of correctly imputed genotypes and genotype accuracy measured as squared Pearson correlation coefficient (r^2) between imputed dosages and typed genotypes. To evaluate the imputation accuracy for different sequencing depth, we randomly sampled reads from the BAM files to generate sequence data with different sequencing depth (1x and 1.5x) using Picard (https://broadinstitute.github.io/picard/). The effects of sample size (number of low coverage sequenced individuals) and minor allele frequency (MAF) on the imputation accuracy were also tested.

Genomic selection

The imputation-based sequence data was used to investigate the performance of genomic prediction in Dezhou Donkey population. Two traits were considered, birth weight (BW) and weaning weight (WW). 594 animals with records on both traits and sequence data were included. The genomic breeding values were estimated using the genomic best linear unbiased prediction (GBLUP) [21] method under single-trait model as well as two-trait model.

Single-trait model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$$

where **y** is the vector of observed phenotypes, **b** is the vector of fixed effects, which include sex effects and year-season effects, **a** is the vector of genomic breeding values with distribution of $N(0, \mathbf{G}\sigma_a^2)$, where σ_a^2 is the additive genetic variance and **G** is the genomic relationship matrix, **X** and **Z** are the incidence matrices for **b** and **a**, respectively, and **e** is the vector of random residuals with distribution of $N(0, \mathbf{I}\sigma_e^2)$.

Two-trait model:

$$\begin{bmatrix}\mathbf{y_1}\\ \mathbf{y_2}\end{bmatrix} = \begin{bmatrix}\mathbf{X_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{X_2}\end{bmatrix}\begin{bmatrix}\mathbf{b_1}\\ \mathbf{b_2}\end{bmatrix} + \begin{bmatrix}\mathbf{Z_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{Z_2}\end{bmatrix}\begin{bmatrix}\mathbf{a_1}\\ \mathbf{a_2}\end{bmatrix} + \begin{bmatrix}\mathbf{e_1}\\ \mathbf{e_2}\end{bmatrix}$$

where the meanings of the vectors and matrices are the same as those in the single-trait model with the subscripts 1 and 2 referring trait 1 and trait 2, respectively. It was assumed that $\begin{bmatrix}\mathbf{a_1}\\ \mathbf{a_2}\end{bmatrix} \sim N(\mathbf{0}, \mathbf{G}\otimes\mathbf{M})$, where $\mathbf{M}=\begin{bmatrix}\boldsymbol{\sigma}^2_{\mathbf{a_1}} & \boldsymbol{\sigma}_{\mathbf{a_{12}}}\\ \boldsymbol{\sigma}_{\mathbf{a_{12}}} & \boldsymbol{\sigma}^2_{\mathbf{a_2}}\end{bmatrix}$ is the variance-covariance matrix of the genomic breeding values of the two traits, and $\begin{bmatrix}\mathbf{e_1}\\ \mathbf{e_2}\end{bmatrix} \sim N(\mathbf{0}, \mathbf{I}\otimes\mathbf{R})$, where $\mathbf{R} = \begin{bmatrix}\boldsymbol{\sigma}^2_{\mathbf{e_1}} & \boldsymbol{\sigma}_{\mathbf{e_{12}}}\\ \boldsymbol{\sigma}_{\mathbf{e_{12}}} & \boldsymbol{\sigma}^2_{\mathbf{e_2}}\end{bmatrix}$ is the residual variance-covariance matrix of the two traits.

Since STITCH provides for each SNP and each individual the imputed genotype (the most likely genotype) as well as the genotype dosages (expected genotypic count, weighted mean of the three genotype values, i.e., 2, 1, and 0, weighted by their corresponding posterior probabilities), the **G** matrix can be constructed using either the imputed genotypes or the genotype dosages. The genotype-based **G** matrix was constructed with the method of VanRaden [22] as follows.

$$\mathbf{G} = \mathbf{W}\mathbf{W}' / \sum 2\,p_j(1-p_j)$$

Here, **W** is the centralized maker genotype matrix with its ijth element equal to

$$w_{ij} = m_{ij} - 2p_j$$

where m_{ij} (= 2,1, or 0) is the original genotype of individual i for SNP j, p_j is the allele frequency of SNP j.

The dosage-based **G** matrix was constructed as follows.

$$\mathbf{G} = \mathbf{D}\mathbf{D}'/s_d$$

Here, **D** is the centralized marker dosage matrix whose elements are zero-centered genotype dosages. s_d is the sum of variances for every column of **D**.

We used both of the two strategies to construct **G**. Different marker densities were also considered when constructing **G** to evaluate the effect of marker density on the performance of genomic prediction.

We used GMAT [23] to construct the **G** matrix and DMU (http://dmu.agrsci.dk) to estimate the variance and covariance components involved in the models and the genomic breeding values (GEBVs).

Cross-validation

In this study, a 12-fold Cross-validation (CV) was applied to assess the accuracy of the genomic selection. The 594 animals were divided into 12 subsets. One of them was taken in turn to be used as validation population and the rest 11 subsets used as training population. The accuracy of genomic prediction for the validation animals was evaluated by $\frac{r_{y_c,GEBV}}{h}$, i.e., the correlation between corrected phenotypic values (y_c) and GEBVs divided by the square root of the heritability. The unbiasedness of predictions was assessed by the regression of y_c on GEBV ($b_{y_c,GEBV}$).

The corrected phenotype for each animal was calculated as the original phenotypic

value corrected for fixed effects (sex and year-season effects) which were estimated using the conventional BLUP model based on the full dataset, i.e., y_c = y – sex effect – year-season effect.

Results

Accuracies of genotype imputation

Comparison of different pipelines

The two genotype imputation pipelines, BaseVar + STITCH and Bcftools + Beagle, were compared using the original sequencing data of the 617 animals with average sequencing depth of 3.5x. It was obvious that the BaseVar + STITCH pipeline was significantly better than the Bcftools + Beagle pipeline (Figure 1), the average genotype accuracy from BaseVar + STITCH was about 6.5 percentage points higher than that from Bcftools + Beagle and the average genotype concordance was about 2.4 percentage points higher. Therefore, the BaseVar + STITCH pipeline was adapted for the subsequent analyses.

The effects of sample size and sequencing depth

We compared the genotype accuracy and genotype concordance for imputation with different sample sizes (200, 400 and 600) and sequencing depths (1x, 1.5x and 3.5x) (Figure 2). In general, as expected, the genotype accuracy and genotype concordance increased with the increase of sample size and sequencing depth. The improvement of imputation accuracy was more obvious when the sample size was

increased from 200 to 400 and the sequencing depth increased from 1x to 1.5x. It should be noted that with sample size of ≥ 400 very high imputation accuracy (genotype accuracy > 0.94, genotype concordance > 0.98) could be achieved even when the sequencing depth was as low as 1x.

The effect of MAF

Figure 3 shows the effect of MAF on imputation accuracy for sample size of 600. For SNPs with MAF < 0.01, the imputation accuracy was greatly affected by MAF and the accuracy increased rapidly with the increase of MAF. However, for SNPs with MAF > 0.01, the imputation accuracy was not affected by MAF.

Variance component estimation

For the single-trait model, the variance components and heritabilities of BW and WW were estimated with the **G** matrix constructed using four levels of marker densities and two strategies (genotype-based **G** matrix and dosage-based **G** matrix). From the original sequence data with an average depth of 3.5x, we obtained 2.3M SNPs after imputation and quality control. We then reduced the marker density by applying LD pruning with three coefficients (r^2: 0.2, 0.4, and 0.8) by PLINK [19], leaving 130K, 220K and 410K SNPs, respectively. The estimates under the four marker densities were very similar (Table 1). The estimates of additive genetic variances and heritabilities based on the dosage-based **G** matrix were all smaller than that based on the genotype-based **G** matrix, although the differences were very small.

For the two-trait model, the variance and co-variance components of the two traits were estimated based on the dosage-based **G** matrix constructed using 410K SNPs (Table 2). The estimates of heritability from the two-trait model (0.627 for birth weight, 0.425 for weaning weight) were higher than that from single-trait model (0.580 for birth weight, 0.330 for weaning weight). The estimate of genetic correlation between birth weight and weaning weight was 0.839.

Accuracy and unbiasedness of GEBVs

The GEBVs for BW and WW were calculated under single-trait model and two-trait model, respectively. For single-trait model, we again considered different **G** matrices constructed with the four levels of marker densities and the two strategies as mentioned above. The accuracies and unbiasedness derived from cross-validation are given in Table 3. Marker density of 410K resulted slightly better results than the other marker densities. The results from the genotype-based **G** matrix and the dosage-based **G** matrix were very similar, with the accuracies from the latter very slightly higher the that from the former. For the two-trait model, only the dosage-based **G** matrix from 410k markers was used. Compared with the results under single-trait model with the same **G** matrix, the two-trait model improved the accuracies by 4 and 5.5 percentage points for BW and WW, respectively, while maintained the same unbiasedness (Table 4).

Discussion

Imputation-based lcWGS data is increasingly being used for genetic analysis of complex traits, such as genome-wide association study and genomic selection/prediction, in human [24-26], plants [27-28] and animals [10, 29-30], and has been proved to be a cost-effective way for genome-wide high-density genotyping, especially for species (such as donkey) for which a SNP array is not available.

Imputation is necessary for lcWGS data due to the high missing rates. Most of the proposed imputation methods designed for lcWGS data infer the gaps between the sparsely mapped reads by leveraging information from a reference panel of haplotypes [31-32]. However, for some animal species there is no such a reference panel available or the size of the panel is not large enough to provide sufficient and reliable information. Some imputation software, like Beagle, can work with or without a haplotype reference panel, however, the imputation accuracy with a haplotype reference panel is much better than that without a haplotype reference panel [12, 33]. Davies et al. (2016) developed a method, STITCH, which was designed specifically for the situation of without a haplotype reference panel, and was proved to yield accuracy comparable with Beagle with a haplotype reference panel [12].

An important step before imputation is to call SNPs from sequence data. GATK [34] and Bcftools[18] are two commonly used software for SNP calling from sequence data. Recently, Liu et al. (2018) develop a SNP calling method, BaseVar [7], which was designed for ultra-low coverage sequencing data. In this study, we compared two pipelines for SNP calling and imputation, i.e., Bcftools + Beagle and BaseVar + STITCH. We demonstrated that BaseVar + STITCH overperformed Bcftools + Beagle

(Figure 1). We showed that in our Dezhou Donkey population, using this pipeline, high imputation accuracy (genotype accuracy > 0.94, genotype concordance > 98%) can be achieved with a sample size of 400 and sequencing depth of 1x (Figure 2). Similar results were also reported by Zhang et al. [10]. In other words, with a sample size of over 400, a sequencing depth of 1x could be sufficient to ensure high imputation accuracy using BaseVar + STITCH.

Using the imputation-based sequence data, we evaluated the performance of genomic selection using GBLUP with respect to four different marker densities (130K, 220K, 410K, and 2.3M), two different **G** matrix construction strategies (genotype-based **G** vs dosage-based **G**), and single-trait vs two-trait models. We found that the prediction accuracy increased slightly when the marker density increased from 130K to 410K. However, it did not further increase when the density increased to 2.3M. The densities of 130K, 220K, and 410K correspond to medium to high density of SNP array, while the 2.3M density corresponds to the density of sequence data. It has been reported that, in the frame of GBLUP, the genomic prediction accuracy could be improved using high density SNP array compared to using medium density array [35-37], however, sequence data could hardly improve the accuracy compared with SNP array [36, 38]. However, sequence data can be meaningful for cross-breed/population genomic selection [39-40]. The accuracy of genomic prediction using the dosage-based **G** matrix is only slightly better than that using the genotype-based **G** matrix. Since the improvement was rather small, we infer this may be due to the high accuracy of imputation.

Noticeable increases in genomic prediction accuracy were observed when using a two-trait model compared with using a single-trait model. It has been long widely proved that multi-trait model can increase the accuracy of breeding value estimation, either by conventional BLUP or by GBLUP [41-43], in particular for traits with high genetic correlation. For the two traits analyzed in this study, birth weight and weaning weight, we obtained an estimated genetic correlation of 0.839, which is very high. This increasement in accuracy with multi-trait model will be particularly beneficial for the situation where the reference population size is limited.

Conclusions

In this study we demonstrated that the pipeline BaseVar + STITCH is a good choice for SNP calling and imputation for low coverage sequence data. Sufficient high imputation accuracy could be achieved for sequence data with sequencing depth as low as 1x, when the size of the sequencing population is over 400. Thus, lcWGS combined with imputation provides a cost-effective way for whole genome high density genotyping and can be applied for large scale genomic selection in farm animals. This is particularly beneficial for those animal species for which a SNP array is not available. In frame of GBLUP, increasing marker density from a density comparable with a high-density SNP array (e.g., 400K) to sequence density with millions of SNPs did not increase the accuracy of genomic selection. Multi-trait model GBLUP improves significantly the accuracy of genomic selection over single-trait model, which would be particularly meaningful for the situation where the reference population size is limited.

Acknowledgements

The authors thank Supercomputing Center in Shandong Agricultural University for technical support.

Authors' contributions

QZ and CN designed the study. CZ, XZ, XJ, HL and SL collected the sample and performed the experiments. CZ and JT analyzed and interpreted the data. CZ, JT, XZ, CN, DW and QZ drafted the manuscript.

Abbreviations

GS: Genomic selection; WGS: Whole genome sequencing; LcWGS: Low coverage whole genome sequencing; MAF: minor allele frequency; BW: Birth weight; WW: weaning weight; GBLUP: Genomic best linear unbiased prediction; GEBV: Genomic breeding value; CV: Cross-validation

Funding

Project for Improved Agricultural Breeding of Shandong Province (2019LZGC011)

Availability of data and materials

All data supporting our findings are included in the manuscript.

Ethics approval and consent to participate

All experimental chickens were maintained, and all the studies were carried out according to the guideline of the experimental animal management of Shandong Agricultural University (SDAUA-2018-018).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4): 1819-29.
2. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet. 2006;123(4): 218-23. doi:10.1111/j.1439-0388.2006.00595.x.
3. Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. Genomic selection in dairy cattle: The USDA experience. Annu Rev Anim Biosci. 2017;5(309-27. doi:10.1146/annurev-animal-021815-111422.
4. Yang AQ, Chen B, Ran ML, Yang GM, Zeng C. The application of genomic selection in pig cross breeding. Yi Chuan. 2020;42(2): 145-52. doi:10.16288/j.yczz.19-253.
5. van Binsbergen R, Calus MP, Bink MC, van Eeuwijk FA, Schrooten C, Veerkamp RF. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2015;47(71. doi:10.1186/s12711-015-0149-x.
6. Moghaddar N, Khansefid M, van der Werf J, Bolormaa S, Duijvesteijn N, Clark SA et al. Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. Genet Sel Evol. 2019;51(1): 72. doi:10.1186/s12711-019-0514-2.
7. Liu S, Huang S, Chen F, Zhao L, Yuan Y, Francis SS et al. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and chinese population history. Cell. 2018;175(2): 347-59. doi:10.1016/j.cell.2018.08.016.
8. Nicod J, Davies RW, Cai N, Hassett C, Goodstadt L, Cosgrove C et al. Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nat Genet.

2016;48(8): 912-8. doi:10.1038/ng.3595.

9. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012;44(6): 631-5. doi:10.1038/ng.2283.

10. Zhang W, Li W, Liu G, Gu L, Ye K, Zhang Y et al. Evaluation for the effect of low-coverage sequencing on genomic selection in large yellow croaker. Aquaculture. 2021;534(

11. Hickey JM. Sequencing millions of animals for genomic selection 2.0. J Anim Breed Genet. 2013;130(5): 331-2. doi:10.1111/jbg.12054.

12. Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence without reference panels. Nat Genet. 2016;48(8): 965-9. doi:10.1038/ng.3594.

13. Hui R, D'Atanasio E, Cassidy LM, Scheib CL, Kivisild T. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. Sci Rep. 2020;10(1): 18542. doi:10.1038/s41598-020-75387-w.

14. Ros-Freixedes R, Gonen S, Gorjanc G, Hickey JM. A method for allocating low-coverage sequencing resources by targeting haplotypes rather than individuals. Genet Sel Evol. 2017;49(1): 78. doi:10.1186/s12711-017-0353-y.

15. Wang C, Li H, Guo Y, Huang J, Sun Y, Min J et al. Donkey genomes provide new insights into domestication and selection for coat color. Nat Commun. 2020;11(1): 6014. doi:10.1038/s41467-020-19813-7.

16. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14): 1754-60. doi:10.1093/bioinformatics/btp324.

17. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al. The Sequence Alignment/Map

format and SAMtools. Bioinformatics. 2009;25(16): 2078-9. doi:10.1093/bioinformatics/btp352.

18. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2). doi:10.1093/gigascience/giab008.

19. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience. 2015;4(7. doi:10.1186/s13742-015-0047-8.

20. Browning BL, Browning SR. Genotype imputation with millions of reference samples. Am J Hum Genet. 2016;98(1): 116-26. doi:10.1016/j.ajhg.2015.11.020.

21. Habier D, Fernando RL, Dekkers JC. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177(4): 2389-97. doi:10.1534/genetics.107.081190.

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

23. Wang D, Tang H, Liu JF, Xu S, Zhang Q, Ning C. Rapid epistatic mixed-model association studies by controlling multiple polygenic effects. Bioinformatics. 2020;36(19): 4833-7. doi:10.1093/bioinformatics/btaa610.

24. Rustagi N, Zhou A, Watkins WS, Gedvilaite E, Wang S, Ramesh N et al. Extremely low-coverage whole genome sequencing in South Asians captures population genomics information. BMC Genomics. 2017;18(1): 396. doi:10.1186/s12864-017-3767-6.

25. Homburger JR, Neben CL, Mishne G, Zhou AY, Kathiresan S, Khera AV. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. Genome Med. 2019;11(1): 74. doi:10.1186/s13073-019-0682-2.

26. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H et al. Extremely low-coverage

sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012;44(6): 631-5. doi:10.1038/ng.2283.

27. Gardner EM, Johnson MG, Ragone D, Wickett NJ, Zerega NJ. Low-coverage, whole-genome sequencing of Artocarpus camansi (Moraceae) for phylogenetic marker development and gene discovery. Appl Plant Sci. 2016;4(7). doi:10.3732/apps.1600017.

28. Zhou C, Duarte T, Silvestre R, Rossel G, Mwanga R, Khan A et al. Insights into population structure of East African sweetpotato cultivars from hybrid assembly of chloroplast genomes. Gates Open Res. 2018;2(41. doi:10.12688/gatesopenres.12856.2.

29. Yang R, Guo X, Zhu D, Bian C, Zhao Y, Tan C et al. Genome-wide association analyses of multiple traits in Duroc pigs using low-coverage. bioRxiv. 2019. doi:10.1101/754671.

30. Xu C, Wu K, Zhang JG, Shen H, Deng HW. Low-, high-coverage, and two-stage DNA sequencing in the design of the genetic association study. Genet Epidemiol. 2017;41(3): 187-97. doi:10.1002/gepi.22015.

31. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021;53(1): 120-6. doi:10.1038/s41588-020-00756-0.

32. Spiliopoulou A, Colombo M, Orchard P, Agakov F, McKeigue P. GeneImp: Fast imputation to large reference panels using genotype likelihoods from ultralow coverage sequencing. Genetics. 2017;206(1): 91-104. doi:10.1534/genetics.117.200063.

33. Zheng C, Boer MP, van Eeuwijk FA. Accurate Genotype Imputation in Multiparental Populations from Low-Coverage Sequence. Genetics. 2018;210(1): 71-82. doi:10.1534/genetics.118.300885.

34. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al. The Genome

Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9): 1297-303. doi:10.1101/gr.107524.110.

35. Boison SA, Utsunomiya A, Santos D, Neves H, Carvalheiro R, Meszaros G et al. Accuracy of genomic predictions in Gyr (Bos indicus) dairy cattle. J Dairy Sci. 2017;100(7): 5479-90. doi:10.3168/jds.2016-11811.

36. Perez-Enciso M, Rincon JC, Legarra A. Sequence- vs. Chip-assisted genomic selection: Accurate biological information is advised. Genet Sel Evol. 2015;47(43. doi:10.1186/s12711-015-0117-5.

37. VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB et al. Genomic imputation and evaluation using high-density Holstein genotypes. J Dairy Sci. 2013;96(1): 668-78. doi:10.3168/jds.2012-5702.

38. Perez-Enciso M. Genomic relationships computed from either next-generation sequence or array SNP data. J Anim Breed Genet. 2014;131(2): 85-96. doi:10.1111/jbg.12074.

39. Druet T, Macleod IM, Hayes BJ. Toward genomic prediction from whole-genome sequence data: Impact of sequencing design on genotype imputation and accuracy of predictions. Heredity (Edinb). 2014;112(1): 39-47. doi:10.1038/hdy.2013.13.

40. van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2014;46(41. doi:10.1186/1297-9686-46-41.

41. Guo G, Zhao F, Wang Y, Zhang Y, Du L, Su G. Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet. 2014;15(30. doi:10.1186/1471-2156-15-30.

42. Jia Y, Jannink JL. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics. 2012;192(4): 1513-22. doi:10.1534/genetics.112.144246.

43. Calus MP, Veerkamp RF. Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol. 2011;43(26. doi:10.1186/1297-9686-43-26.

Figures

Figure 1 Comparison of imputation accuracy using two imputation pipelines

BaseVar + STITCH: Calling SNPs using BaseVar and imputing missing genotypes using STITCH

Bcftools + Beagle: Calling SNPs using Bcftools and imputing missing genotypes using Beagle

a, b and c: genotype accuracy for chromosomes 1, 19, and 30, respectively;

d, e and f: genotype concordance for chromosomes 1, 19 and 30, respectively

Figure 2 The effects of sample size and sequencing depth on imputation accuracy using the pipeline BaseVar + STITCH

a, b and c: genotype accuracy for chromosomes 1, 19, and 30, respectively;

d, e and f: genotype concordance for chromosomes 1, 19, and 30, respectively

Figure 3 The effect of minor allele frequency on imputation accuracy using the pipeline BaseVar + STITCH

a, b and c: genotype accuracy for chromosomes 1, 19, and 30, respectively;

d, e and f: genotype concordance for chromosomes 1, 19, and 30, respectively

Table 1 Estimates of variance components and heritabilities based on single-trait model for birth weight (BW) and weaning weight (WW)in Dezhou Donkey population

Variance	130K		220K		410K		2.3M	
	G(g)	G(d)	G(g)	G(d)	G(g)	G(d)	G(g)	G(d)
				BW				
σ_a^2	10.266	10.100	10.424	10.253	10.750	10.581	10.166	10.011
σ_e^2	8.024	8.030	7.904	7.912	7.656	7.664	8.154	8.159
h^2	0.561	0.557	0.569	0.564	0.584	0.580	0.555	0.551
				WW				
σ_a^2	68.419	68.223	69.136	68.919	69.866	69.670	62.042	61.083
σ_e^2	139.977	140.874	140.659	141.777	141.046	141.328	146.029	146.074
h^2	0.328	0.326	0.330	0.327	0.331	0.330	0.298	0.295

σ_a^2: additive genetic variance; σ_e^2: residual variance; h^2: heritability; G(g): genotype-based G matrix; G(d): dosage-based G matrix

Table 2 Estimates of variance components, heritabilities, and genetic correlation for birth weight (BW) and weaning weight (WW) under a two-trait model

Variance	Two-trait model[a]	
	BW	WW
σ_a^2	11.769	90.728
σ_e^2	7.016	122.553
h^2	0.627	0.425
r_g	0.839	

a: using dosage-based G matrix constructed with 410K SNPs

σ_a^2: additive genetic variance; σ_e^2: residual variance; h^2: heritability; r_g: genetic correlation;

Table 3 Accuracies (±*SEs*) and unbiasedness (±*SEs*) of genomic selection under a single-trait model for birth weight (BW) and weaning weight (WW)

Marker density	Genotype-based **G** matrix		Dosage-based **G** matrix	
	Accuracy	Unbiasedness	Accuracy	Unbiasedness
		BW		
130K	0.368±0.042	0.983±0.141	0.370±0.041	0.983±0.141
220K	0.370±0.042	0.984±0.139	0.371±0.040	0.984±0.140
410K	0.376±0.041	0.990±0.135	0.378±0.040	0.990±0.136
2.3M	0.371±0.044	0.980±0.140	0.372±0.042	0.981±0.140
		WW		
130K	0.399±0.030	1.198±0.209	0.400±0.030	1.199±0.210
220K	0.402±0.030	1.191±0.213	0.403±0.029	1.191±0.209
410K	0.409±0.029	1.187±0.214	0.410±0.029	1.187±0.214
2.3M	0.403±0.032	1.212±0.221	0.405±0.031	1.212±0.221

Table 4 Accuracies (±*SEs*) and unbiasedness (±*SEs*) of genomic selection for birth weight (BW) and weaning weight (WW) under a single- and a two-trait model using dosage-based **G** matrix constructed with 410K SNPs

Model	BW		WW	
	Accuracy	Unbiasedness	Accuracy	Unbiasedness
Single-trait model	0.378±0.041	0.990±0.136	0.410±0.029	1.187±0.214
Two-trait model	0.417±0.039	0.991±0.113	0.465±0.039	1.182±0.153

Figures

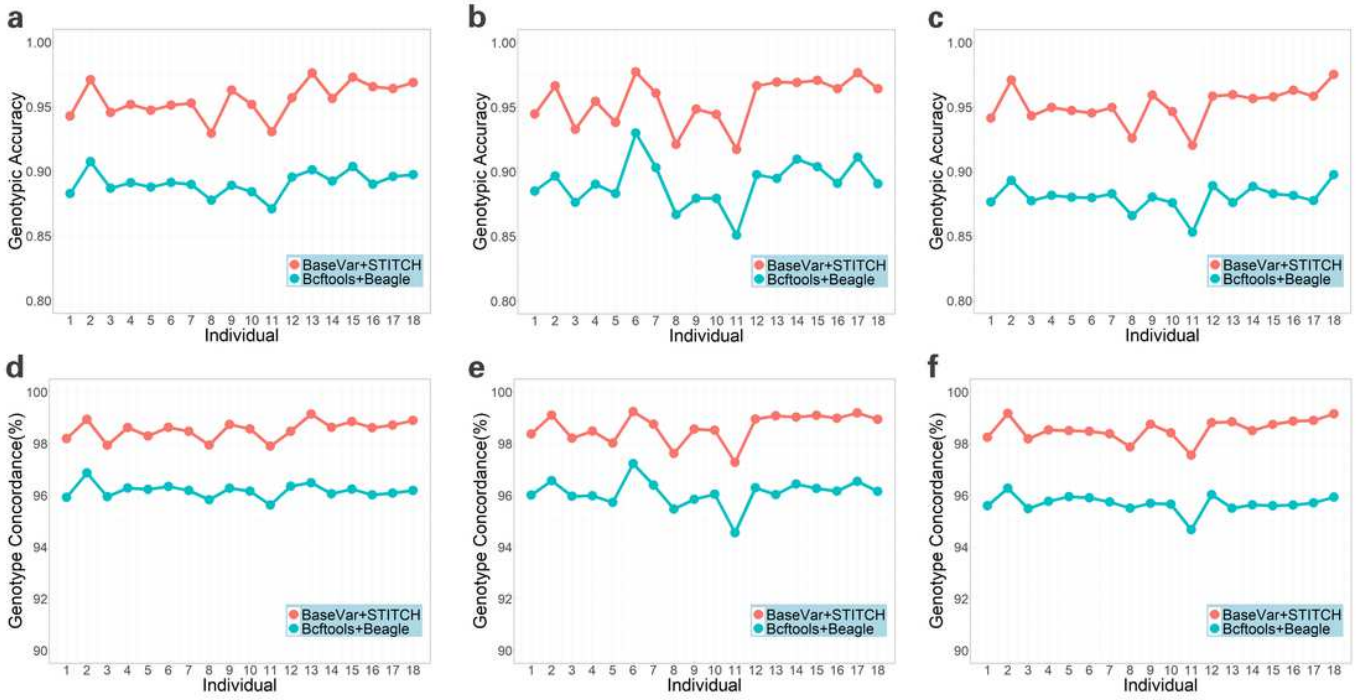

Figure 1

Comparison of imputation accuracy using two imputation pipelines BaseVar + STITCH: Calling SNPs using BaseVar and imputing missing genotypes using STITCH Bcftools + Beagle: Calling SNPs using Bcftools and imputing missing genotypes using Beagle a, b and c: genotype accuracy for chromosomes 1, 19, and 30, respectively; d, e and f: genotype concordance for chromosomes 1, 19 and 30, respectively

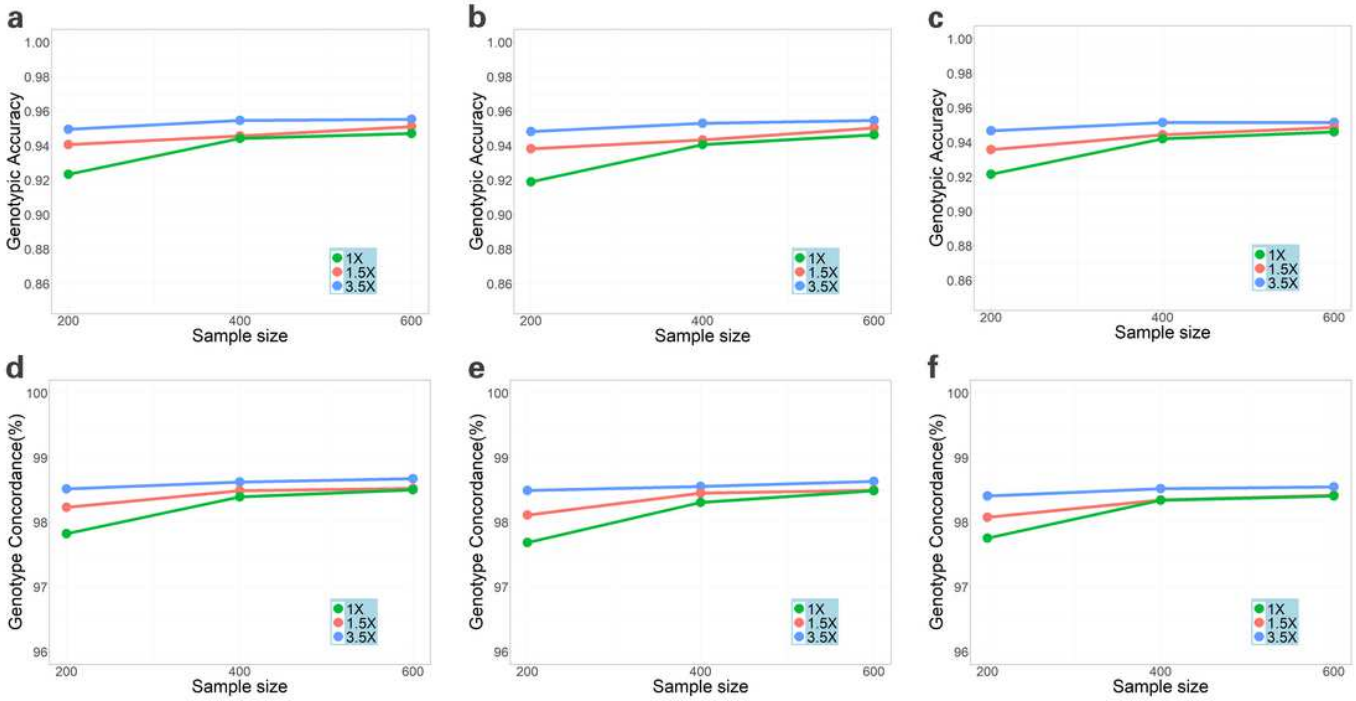

Figure 2

The effects of sample size and sequencing depth on imputation accuracy using the pipeline BaseVar + STITCH a, b and c: genotype accuracy for chromosomes 1, 19, and 30, respectively; d, e and f: genotype concordance for chromosomes 1, 19, and 30, respectively

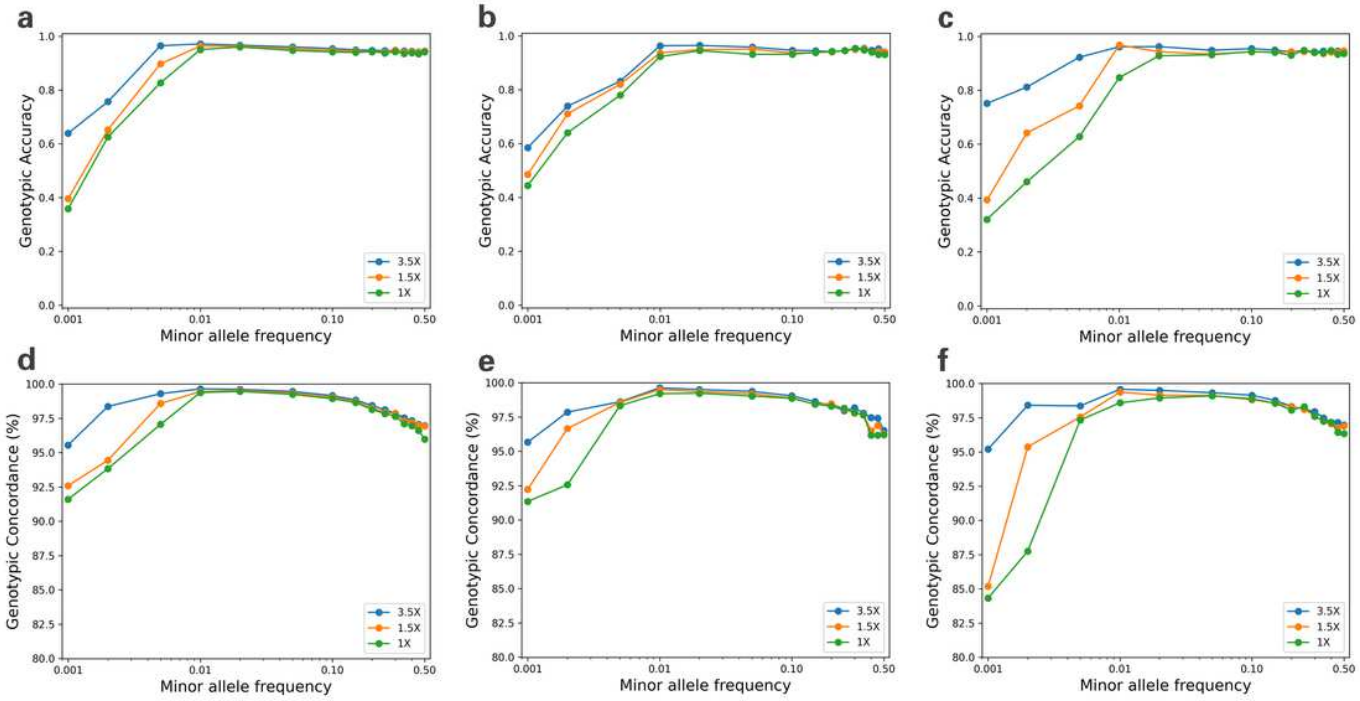

Figure 3

The effect of minor allele frequency on imputation accuracy using the pipeline BaseVar + STITCH a, b and c: genotype accuracy for chromosomes 1, 19, and 30, respectively; d, e and f: genotype concordance for chromosomes 1, 19, and 30, respectively