# Low-coverage Whole-genome Sequencing Accelerates Decoding of the Genetic Architecture of Economic Traits in Domestic Pigs

## MGI DNBSEQ-G400 sequencing platform enables genetic structure analysis of important economic traits in pigs

Based on years of research, Prof. Hu Xiaoxiang from China Agricultural University and Prof. Wu Zhenfang from South China Agricultural University developed a low-coverage whole-genome sequencing analysis method for Duroc pigs based on DNBSEQ-G400 sequencing platform, and analyzed the genetic architecture of important economic traits in pigs by high-throughput sequencing technology.

The research was published in *GigaScience* in 2021 entitled "Accelerated deciphering of the genetic architecture of agricultural economic traits in pigs using a low-coverage whole-genome sequencing strategy "[1].

Recommended application: Molecular breeding

Recommended model: DNBSEQ-G400RS

- ### Data output is efficient and high-quality

  DNBSEQ™ sequencing technology has significant features such as high accuracy, low repeat sequence rate, and low index hopping rate, and its low-coverage whole-genome sequencing (LcWGS) genotyping is more accurate compared to other methods[1].

- ### Offer a complete product combination for experimental procedures

  Based on independently developed automatic solutions and analysis software, MGI provides a complete set of product combination from sample to result output.

# Background

As the world's largest pork consumer and pig producer, China has introduced up to 90% of pig breeds, and in the face of the grim status quo of "neck stucking" in pig industry, it is crucial to promote the breeding of good strains. In recent years, with the rapid development of high-throughput sequencing technology, whole genome sequencing, functional genomics and bioinformatics of species have been developed rapidly, however, the high cost of whole genome sequencing restricts its commercialization and industrial transformation. The low-coverage whole-genome sequencing (lcWGS) offers new options for strain selection in the livestock and food industries[2].

LcWGS refers to performing sequencing at a coverage of approximately 1× or less[3, 4] and then restoring the lost genotypes/loci by imputation to ensure that all individuals have a set of genotypes sharing the variation. This method has been used for genome-wide association studies (GWAS) and genome selection/prediction in humans and some animal species, and has proven to be a viable alternative to high-coverage sequencing[5]. Overall, lcWGS is a new low-cost sequencing marker approach. It has been shown that lcWGS of a large number of individuals is more accurate than sequencing fewer individuals at higher coverage rate for the assessment of population parameters[6].

## Study description

The research team collected 2869 Duroc boars from the same breeding farm and performed genome library preparation based on Tn5 transposase approach. The libraries were sequenced with an average coverage of 0.73× on DNBSEQ-G400 sequencing platform. The BaseVar-Stitch process was used for reference panel preparation and genotype imputation. And three different genotyping methods (high-coverage sequencing, SNP Array, Fluidigm Integrated Fluidic Circuit (IFC) direct genotyping) were used to evaluate accuracy and compare the effect of sample size and sequencing coverage on accuracy under different parameters. The whole-genome association studies and genetic architecture analysis of 21 economically important traits of pigs were performed by low-coverage sequencing, and the effect of artificial selection on genome structure changes during the selection and breeding of Duroc pigs was systematically analyzed (Figure 1).

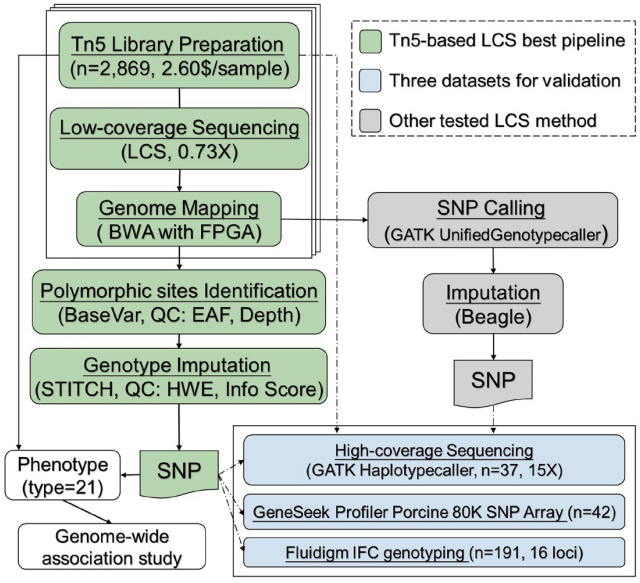


**Figure 1.** The flow chart of low-coverage whole-genome sequencing.

## Materials and Methods

### A.Sample collection

In this study, the ear tissues of 2869 Duroc boars from the same breeding farm were collected and genomic DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen 69506).

### B.Library preparation and sequencing

50 ng of gDNA was transferred and fragmented by Tn5 transposase. PCR amplification was performed using KAPA HiFi HotStart Ready Mix (Roche). The amplification primers were designed based on the MGI sequencer and the reverse primers contain 96 different barcode sequences to distinguish individual libraries. After the library was prepared into DNA nanoballs (DNBs), paired-end 100 bp (PE100) sequencing was performed on DNBSEQ-G400 or BGISEQ-500 sequencing platform.

### C.Data analysis

GTX-ALIGN was applied to map sequencing reads of low-coverage samples to the Sscrofa11.1 reference genome, and the indel realignment and base quality score recalibration modules in GATK were used to realign reads against the indel candidate loci and recalibrate the base quality. Population genetics analysis was performed using the vcfTools program and BioMart tool. A mixed linear model method was applied to perform genome-wide association analysis based on tagging SNPs.

| Sample collection | Library preparation and sequencing | Bioinformatics analysis | Result analysis |
|---|---|---|---|
| 2869 Duroc boars from the same breeding farm | Tn5-based WGS library preparation<br>  <br>DNBSEQ-G400 genetic sequencer | GTX-ALIGN<br>vcfTools BioMart | Comparison of genotyping effects on different platforms,<br>Performance evaluation of LcWGS-based analysis process,<br>Analysis of genetic architecture of economically important traits |

# Results

## Comparison of genotyping effects using different platforms

The MGI sequencing platform showed the advantages of low duplication rate, high data output rate and low data loss rate in genotyping (Table 1). Therefore, the follow-up study was conducted based on DNBSEQ-G400 sequencing platform.

| | illumina | MGI |
|---|---|---|
| platform | X-ten | MGISEQ-2000 |
| Samples/lane | 96 | 84 |
| Bases number | ~136G | ~151G |
| Reads number | ~455M | ~755M |
| Good index bases | ~109G | ~138G |
| Total coverage | 98.51% | 98.55% |
| Ave coverage | 29.03% | 42.56% |
| Ave depth | 0.40 | 0.63 |
| Mapping rate | 99.28% | 97.86% |
| PCR duplication | 12.64% | 4.27% |

**Table 1.** Comparison of genotyping results between Illumina sequencing platform and MGI sequencing platform. The DNBSEQ G400 is also known as MGISEQ-2000 in China and some other markets.

## Performance evaluation of LcWGS-based BaseVar-STITCH pipeline

Low-coverage sequencing data processed by BaseVar-STITCH pipeline resulted in higher accuracy genotypes ($R^2$=0.919, GC=0.970) compared to 15x whole-genome sequencing results, and the accuracy exceeded the results of the GATK-Beagle-based analysis ($R^2$=0.484, GC=0.709) (Figure 2A). The BaseVar-STITCH results showed higher GC consistency and $R^2$ values ($R^2$ = 0.997, GC = 0.990) compared to GGP-80 data (SNP Array) (Figure 2B). The average GC for Fluidigmn-based genotyping (16 loci, 191 individuals) was 0.991, close to the BaseVar STITCH data.

In summary, the BaseVar-STITCH pipeline is a new analytical method suitable for variant detection and gene imputation for LcWGS strategy. When the sample size was >500, there was almost no effect on the results when 0.5× coverage was used for STITCH analysis. When the sample size was increased to 1985, the results at a sequencing coverage of 0.1× were consistent with those at the 0.5× coverage (Figure 2C and D). Overall, the accuracy of results

improves significantly with increased sequencing coverage or sample size, and the high confidence of genotyping in the study can be guaranteed when the total sequencing coverage of individual locus is >200×.

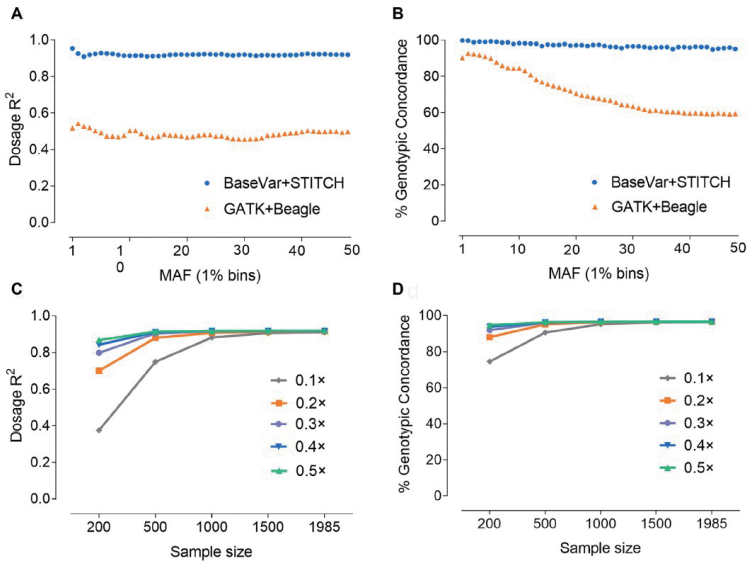


**Figure 2.** Performance of BaseVar-STITCH in different minor allele frequencies (MAFs) and sample sizes. (Note: Correlation between genotypes and imputed dosages ($R^2$) and genotypic concordance (GC) were calculated to evaluate genotype accuracy)

## Genetic architecture of economically important traits in pigs

The GWAS and genetic architecture of 21 economically important traits in pigs were analyzed (Figure 3), and the effects of artificial selection on the genome structure of pigs during the breeding process were systematically analyzed. For the identified major QTL loci, the major candidate genes *ABCD4* associated with total teat number and *HMGA1* associated with back fat thickness trait were further identified by high-density marker method; For traits with high heritability but controlled by multiple minor genes, numerous minor genes involved in important neural pathways affecting feeding behavior were identified, and the classical "infinitesimal model" for quantitative traits was further validated, which provided guidance for the next generation of molecular breeding.

In addition, this study also found that traits related to growth showed QTL fixation and loss of heritability changes due to the long-term artificial selection of this population, indicating that the parental population has achieved remarkable results in the selection of growth traits in the previous cross breeding.

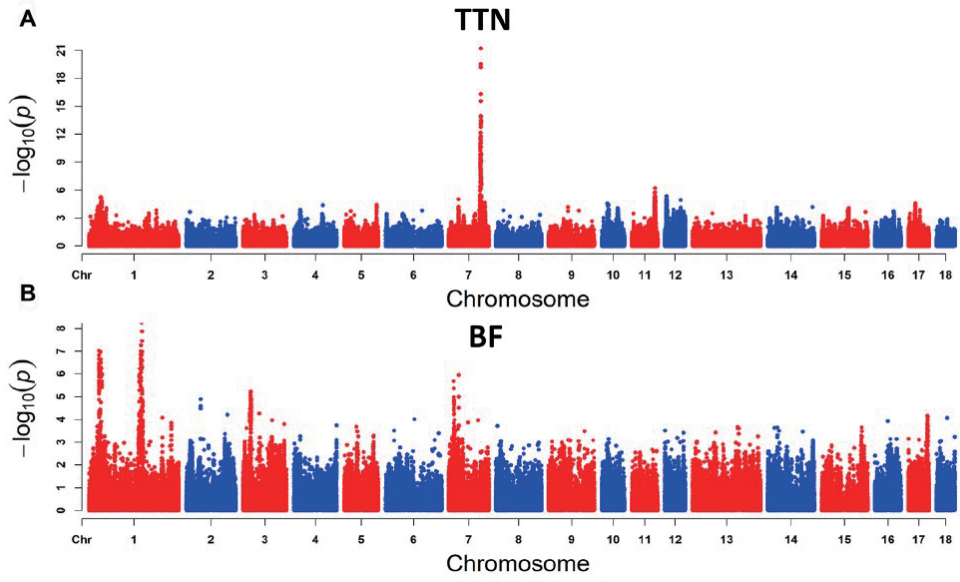


**Figure 3.** Manhattan plots of total teat number (TTN) and backfat thickness (BF). A and B described the association signals of TTN and BF on the whole-genome.

## Summary

This study invented the first BaseVar-Stitch genotyping pipeline using low-coverage sequencing to obtain a high-density SNP marker set (11.7M) for the largest population (2869) of Duroc pigs with a very low cost. The genotyping accuracy by three different methods exceeded 99%, which demonstrated the superiority of the large-sample, low-coverage sequencing strategy created in this study over the traditional high-depth sequencing based on small samples. When combined with MGI's automation solutions for experimental processes and algorithm optimization, the labor cost and experimental period can be dropped significantly.

Meanwhile, further studies on the genetic parameters and genetic architecture of growth, carcass and feeding traits were carried out by GWAS. For major QTL associated with pig teat number, the candidate major genes *ABCD4* and *HMGA1* were identified using a one-step method with high-density markers. For traits with high heritability but controlled by minor genes, a large number of minor genes involved in important neural pathways affecting feeding behavior were identified, and the classical "infinitesimal model" of quantitative genetics was further confirmed, which offered a theoretical basis for molecular breeding. The traits related to growth traits in this population showed QTL fixation and loss of heritability changes, suggesting the remarkable success of the growth-related selection and breeding in this parental population in the previous cross breeding.

This study adopted the low-coverage sequencing technology to fill in the high efficiency and high-quality SNP imputation method in livestock

and poultry species without a good reference panel. The researchers proposed a solution to the important "neck stucking" problem in breeding and demonstrated its potential value in functional gene mapping and genetic structure analysis of important economic traits. Furthermore, the method offered a new genetic analysis strategy and breeding guidance with completely independent intellectual property rights to promote the Seed Industry Revitalization Action Plan in China's 14th Five-Year Plan.


Gene Sequencer DNBSEQ-G400RS

## Reference


1. Yang, R. *et al.* Accelerated deciphering of the genetic architecture of agricultural economic traits in pigs using a low-coverage whole-genome sequencing strategy. *GigaScience* 10,doi:10.1093/gigascience/giab048 (2021).
2. Casellas, J., Martín de Hijas-Villalba, M.,Vázquez-Gómez, M. & Id-Lahoucine, S. Low-coverage whole-genome sequencing in livestock species for individual traceability and parentage testing. *Livestock Science* 251, 104629,doi:https://doi.org/10.1016/j.livsci.2021.104629 (2021).
3. Nosková, A. *et al.* Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in Swiss Large White pigs. *BMC genomics* 22, 290,doi:10.1186/s12864-021-07610-5 (2021).
4. Buckley, R. M. *et al.* Best practices for analyzing imputed genotypes from low-pass sequencing in dogs. *Mammalian genome : official journal of the International Mammalian Genome Society* 33, 213-229, doi:10.1007/s00335-021-09914-z (2022).
5. Zhao, C. *et al.* Towards a Cost-Effective Implementation of Genomic Prediction Based on Low Coverage Whole Genome Sequencing in Dezhou Donkey. *Frontiers in genetics* 12, 728764,doi:10.3389/fgene.2021.728764 (2021).
6. Lou, R. N., Jacobs, A., Wilder, A. P. & Therkildsen, N. O. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular ecology* 30, 5966-5993,doi:10.1111/mec.16077 (2021).

## Recommended Ordering Information

| Category | Product | Cat. NO. |
|---|---|---|
| Instruments | Genetic Sequencer DNBSEQ-G400RS | 900-000170-00 |
| | MGISP-100RS Automated Sample Preparation System | 900-000206-00 |
| | MGISP-960RS Automated Sample Preparation System | 900-000146-00 |
| Software | MegaBOLT Bioinformatics analysis accelerator | 900-000555-00 |
| Library Prep | MGIEasy Universal DNA Library Prep Set (16 RXN) | 1000006985 |
| Sequencing Reagents | DNBSEQ-G400RS High-throughput Sequencing Set (FCL PE150) | 1000016952 |


MGI Tech Co.,Ltd | Stock abbreviation: MGI
Stock code: 688114

Building 11, Beishan Industrial Zone, Yantian District, Shenzhen, CHINA, 518083





Version: August 2023



+86-4000-688-114
en.mgi-tech.com
MGI-service@mgi-tech.com



Author: Li Yanping, Sui Jiayuan
Editor-in-Charge: Qiwei Wang
Reviewer: Yao Jiang


1. For StandardMPS and CoolMPS: Unless otherwise informed, StandardMPS and CoolMPS sequencing reagents, and sequencers for use with such reagents are not available in Germany, Spain, UK, Sweden, Belgium, Italy, Finland, Czech Republic, Switzerland, Portugal, Austria and Romania. Unless otherwise informed, StandardMPS sequencing reagents, and sequencers for use with such reagents are not available in Hong Kong. No purchase orders for StandardMPS products will be accepted in the USA until after January 1, 2023.
2. For HotMPS sequencers: This sequencer is only available in selected countries, and its software has been specially configured to be used in conjunction with MGI's HotMPS sequencing reagents exclusively.
3. For HotMPS reagents: This sequencing reagent is only available in selected countries.