Enhancing Breeding through Genomic Selection

Application solutions based on MGI sequencing platforms

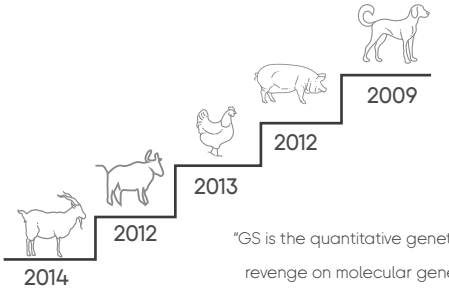
Development of genomic selection

Genomic selection (GS) is a new generation of molecular breeding technology. It uses high-density genome-wide markers to calculate the Genomic Estimated Breeding Value (GEBV) of individuals, which contributes to the early selection, shorter generation interval, and faster breeding process. Because of the high accuracy and increased trait selection rate, GS have been used as an important approach to improve genetic traits of livestock and crops currently.

Best line selection based on traits	Best line selection based on molecular marker-assisted selection	Best line selection with multiple traits based on genome-wide markers
Phenotypic selection (PS)	**Mark assisted selection (MAS)**	**Genomic selection (GS)**
1	2	3
• Longer generation interval needed • Quantitative traits greatly affected by environmental factors	• Slow breeding process • Only selected genetic variations are targeted • Difficulty in mapping quantitative trait loci	• High selection accuracy at early stage • Reduced generation interval and accelerated genetic progress • A substantial increase in population genetic improvement rate and production efficiency

Genomic selection was first described in 2001 by Meuwissen and colleagues (1). It was further demonstrated with the huge potential in dairy cattle breeding in 2006 (2). Since then, the technology has become a research hotspot in the molecular breeding of livestock and crops and a competition focus among multinational companies.

Advances in the sequencing technology have led to the substantial reduction in the cost of genome sequencing. Genome selection based on the sequencing technology become increasingly popular in the breeding of agricultural animals such as dairy cattle, pigs, sheep, chickens, ducks, etc. This technology is also being frequently used in important agricultural crops and forest tree breeding.



"GS is the quantitative geneticists revenge on molecular genetics"
——A.Archibald

References

[1] Meuwissen T H E, Hayes B J, Goddard M E. Prediction of total genetic value using genome-wide dense marker maps[J]. **Genetics**, 2001, 157(4): 1819-1829.

[2] Schaeffer L R. Strategy for applying genome - wide selection in dairy cattle[J]. **Journal of animal Breeding and genetics**, 2006, 123(4): 218-223.

Limitations of DNA arrays in genomic selection applications

Dependency on prede-signed SNP information

1

- Only known SNPs are targeted
- Unable to discover new SNPs

Missing low frequency mutation information

2

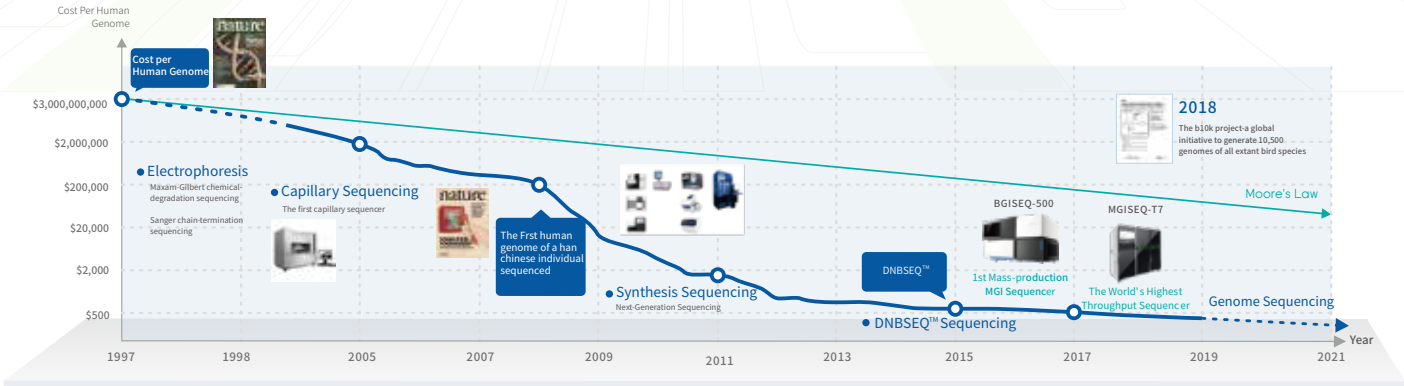
- The screened SNP sites mostly contain polymorphism information

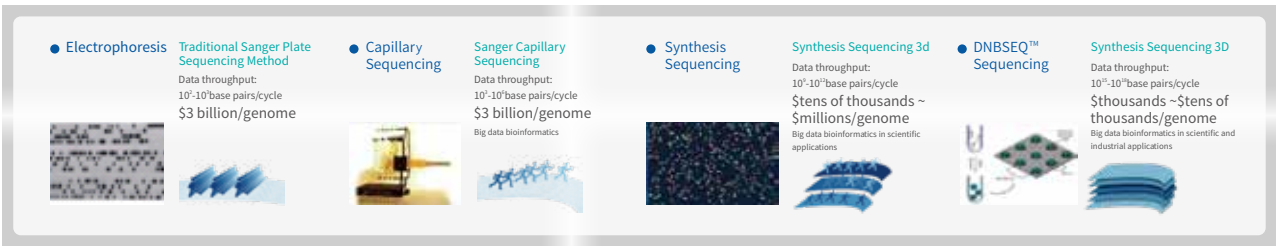
Low compatibility

3

- Designed for commercial species
- Not applicable to world-wide species

Continuous reduction in sequencing costs leads to widespread use of genomic selection





MGI solutions for genomic selection

DNBSEQ™ platform provides high-quality and cost-effective sequencing scheme for agricultural genome research.

Automated sample preparation system > **High-throughput sequencing platform** > **Molecular breeding analysis system**

MGISP-960

- ☐ Automated nucleic acid extraction
- ☐ Automated library preparation

DNBSEQ-G50 DNBSEQ-G400 DNBSEQ-T7

- ☐ Targeted sequencing
- ☐ Restriction-site Associated DNA (RAD) sequencing
- ☐ Low-coverage whole-genome sequencing

Computing workstation Server

- ☐ Data storage
- ☐ Data analysis
- ☐ Report output

High-throughput sequencing platform

MGI provides various sequencing platforms with different throughputs, such as DNBSEQ-G50, DNBSEQ-G400*, and DNBSEQ-T7*, to meet the needs of enterprises, research institutes and government agencies for different sample sizes and detection speeds in molecular breeding applications. For example, when a genotyping approach based on low-coverage whole-genome sequencing (lcWGS) of a large cohort was used as genomic selection for pigs, different MGI sequencing platforms were compared in terms of annual detection volume and operation cycles as shown in the table below.

Product model	DNBSEQ-G50	DNBSEQ-G400	DNBSEQ-T7
Features	Flexible	Adaptive	Ultra-high throughput
Flow cell type	FCS & FCL	FCS & FCL	FC
Flow cell / run	1	1-2	1-4
Effective read* / Flow cell	100 M/500 M	550 M/1500-1800 M	5000 M
Recommended reads	PE100	PE100	PE100
Max. throughput / day	90 Gb	360 Gb	4 Tb
Number of samples / day	48	192	2,000
Annual throughput **	7,200 samples	28,800 samples	300,000 samples

* The maximum number of effective reads and Data output are based on the sequencing of an internal standard library. Actual output may vary depending on sample type and library preparation method.

** The annual throughput is calculated based on 3 rounds per week and 50 weeks per year

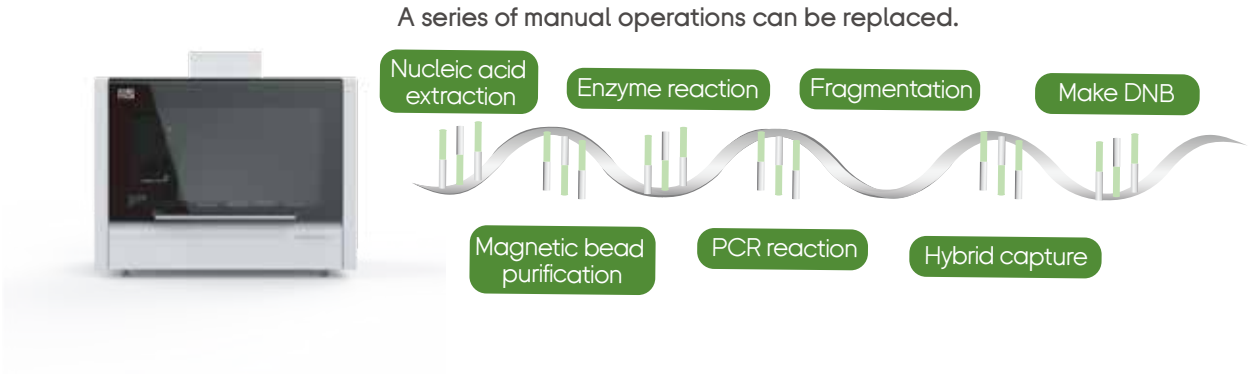
*** Take 0.5x WGS (pig, 3G) as an example

Automated sample preparation system

MGISP-960 High-throughput automated sample preparation system is a flexible and efficient fully-automated workstation equipped with 96-channel pipettes. With this fully automatic operation design, it can realize unmanned process for a series of operations such as nucleic acid extraction and library construction. It can also be customized according to customer needs.

Nucleic acid extraction and library construction can be completed within 5 hours.

With MGIEasy Magnetic Beads Genomic DNA Extraction kit and MGIEasy Fast PCR-FREE FS DNA Library Prep set, DNA extraction from 96 samples can be completed within 2.5 hours, and the PCR-free WGS library construction can be completed within 2.5 hours. That means the nucleic acid extraction and library construction can be completed within 5 hours, and the daily throughput can reach up to 384 samples.



Molecular Breeding Analysis System

In the analysis of genomic selection, the genotype and phenotype data are large in volume which require huge storage and complex analysis process. As such, challenges such as data storage, analysis and management need to be addressed. Molecular breeding analysis system helps molecular breeding advance to precision breeding.

All-in-one computing workstation

- Data storage
- Data analysis
- Data management

Data storage	Data analysis	Report output
• Sequencing data • Phenotypic data • Flow cell data • Analyzed data • Sample-related information	• Data quality control • Variant detection • Genotype imputation • Filtering SNP • Breeding value calculation • Genomic mating	• Reporting, ranking filtering, and best individual selection • Serve breeding industry by assisting in selection

Cases of MGI platforms in genomic selection

Case 1: Low-depth whole-genome sequencing ——Developing a total molecular breeding solution for livestock and poultry based on MGISEQ-2000 (DNBSEQ-G400)

Research proposal

The research teams from two China universities collaborated to develop new methods for sequencing large cohorts without large reference panels. 2869 Duroc boars from the same breeding farm were sampled and the genome library was constructed using the self-optimized protocol. Low-coverage whole genome sequencing was performed on a MGISEQ-2000 (DNBSEQ-G400) platform at a mean depth of 0.73×. The BaseVar-Stitch process is chosen for Reference Panel construction and genotype imputation. At the same time, three different genotyping methods (SNP chip, high-depth sequencing, Fluidigm genotyping) were used to evaluate the accuracy of low-depth data, and different parameters were used to evaluate the impact of sample size and sequencing depth on the accuracy of results (Figure 1).

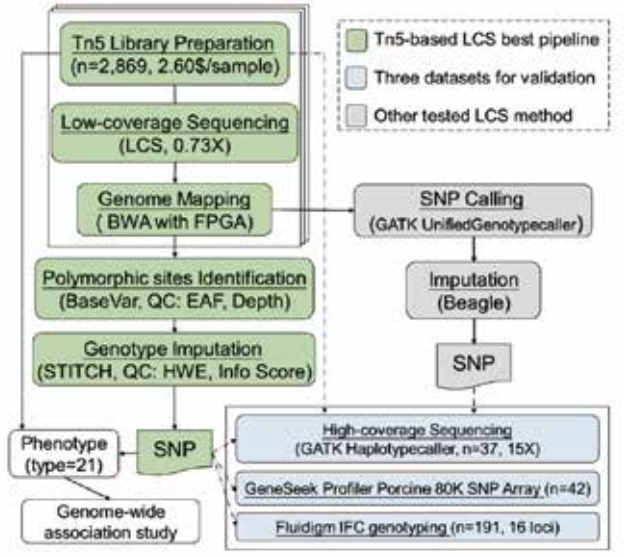


Figure 1. Low-coverage whole-genome sequencing protocol design

Results

1. Performance evaluation of BaseVar-STITCH analysis process based on lcWGS

Highly accurate genotypes were obtained using the BaseVar-STITCH pipeline compared with the high-depth sequencing result (R^2 = 0.919 and GC = 0.970) which exceeded the method using GATK-Beagle (R^2 = 0.484 and GC = 0.709) (Figure 2A). Moreover, the BaseVar-STITCH results showed even higher GC concordance and R^2 values compared with the SNP chip GGP-80 data (R^2 = 0.997 and GC = 0.990)(Figure 2B). Furthermore, direct genotyping (16 loci, 191 individuals) was carried out using the Fluidigm dynamic array IFC. The mean GC was 0.991 compared with the BaseVar-STITCH data, which is as high as the aforementioned results. Taken together, these results suggest that BaseVar-STITCH pipeline is a suitable variant discovery and imputation method for the lcWGS strategy.

2. The sequencing depth of lcWGS can be as low as 0.1×

For the 0.5× coverage using STITCH, a sample size >500 had little effect on performance. At a 0.1× downsampled coverage, increasing the sample size to 1,985 led to a substantially improved performance (Figure 2 C and D). In general, the total sequencing depth (population category) for 1 locus >200× was shown to guarantee the credibility of genotyping within the scope of this study, although the results consistently improved as sequencing depth/sample size increased.

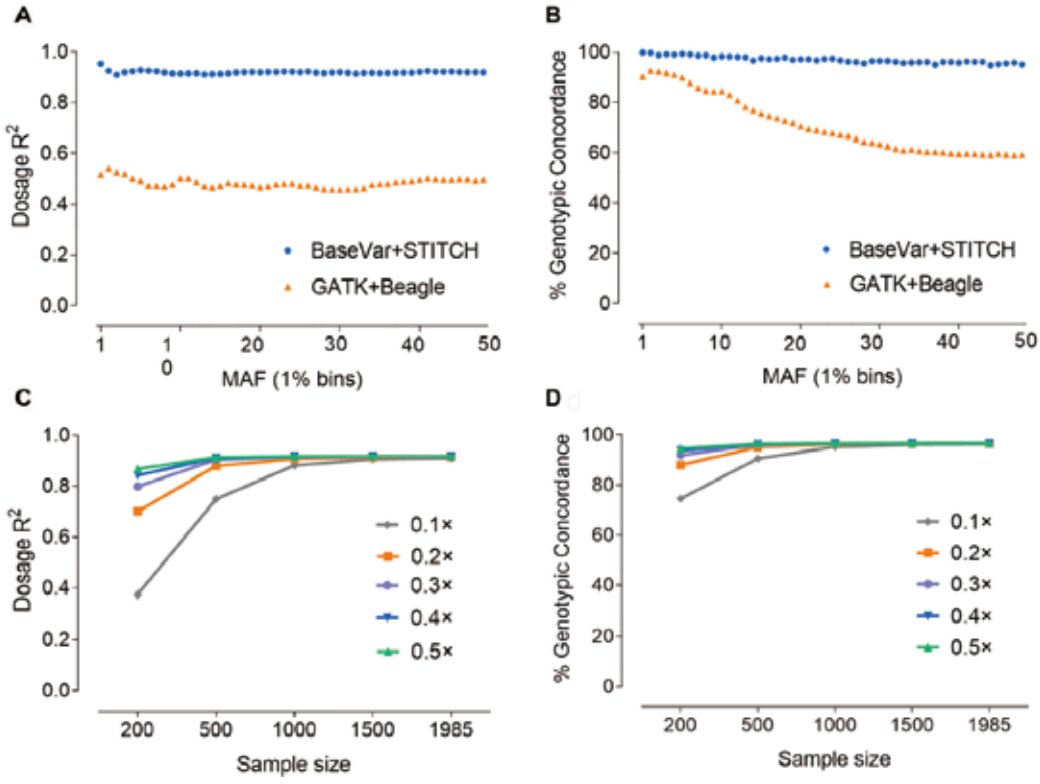


Figure 2. Performance of BaseVar-STITCH on different minor allele frequencies (MAFs) and sample sizes

NOTE: Calculate the correlation (R^2) and genotype concordance (GC) between genotype and estimated dose to evaluate genotype accuracy

Conclusions

In this study, the BaseVar-Stitch genotyping pipeline based on low-depth sequencing was established for the first time. The high-density SNP marker set (11.7M) of by far the largest cohort of 2869 Duroc pigs was obtained at a very low cost. The genotyping accuracy assessed by different methods is over 99%, which proves that the imputation strategy using a large cohort without a good reference panel is significantly advanced compared to the traditional high-depth data imputation based on small sample size.

Reference

Yang R, Guo X, Zhu D, et al. Accelerated deciphering of the genetic architecture of agricultural economic traits in pigs using a low-coverage whole-genome sequencing strategy [J]. **GigaScience**, 2021, 10(7): giab048.

Case 2: Targeted Sequencing ——Development of grouper breeding panel based on customized ATOPlex multiplex PCR and MGISEQ-2000 (DNBSEQ-G400)

Experimental design

By using MGI's customized ATOPlex multiplex PCR platform, BGI Marine customized a panel for grouper breeding based on 741 independent SNPs related to grouper body weight and ammonia tolerance. The specific panel design was as follows:

Table 1. Details of customized panel designed for grouper breeding

Product name	Customized grouper breeding panel
Amplicons number	740+ plex
Amplicon size	100-200 bp
Variant detection	741 SNPs
Detection principle	Multiplex PCR + high-throughput sequencing
Sample type	gDNA
DNA input	1~10 ng DNA
Recommended reads amount	1.0 M reads/sample

Results

1. The panel has good performance, including >99% mapping rate, >98% on-target rate and >95% uniformity (Figure 3).
2. The sequencing depth of all loci is over 500X, with high data homogeneity (Figure 4).
3. Consistent sequencing depth in regions with different GC contents (Figure 5).
4. The accuracy of the panel's prediction of grouper body weight is over 80%, and the accuracy of ammonia tolerance is over 95%. The customized panel has a detection rate of over 94% for different grouper species.

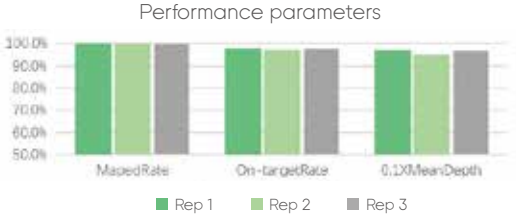


Figure 3. Performance parameters

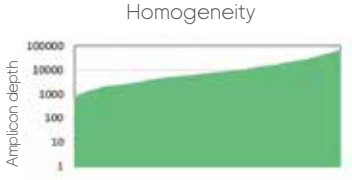


Figure 4. Homogeneity of output data

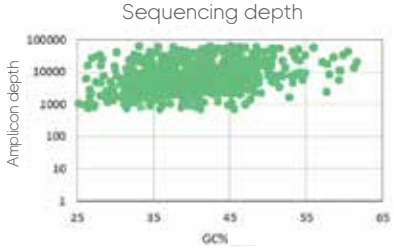


Figure 5. Sequencing depth of different GC content

Case 3: Targeted Sequencing
——Development of porcine GBTS liquid chip series products based on MGI sequencing platform

Experimental design

Based on the scientific research achievements from Shandong Agricultural University/China Agricultural University and the published results of pig genome research in recent years, Molbreeding Biotechnology, a genotyping solution provider, developed 50K, 40K, 10K, 1K porcine liquid-based chips. Different from the solid phase-based DNA arrays, the liquid chip based on the MGI sequencing platform has the advantages of flexible panel design and high detection throughput. The developed liquid chips can be used in various applications for the popular pig breeds (Yorkshire, Landrace and Duroc), such as genomic selection, pedigree identification, genetic diversity analysis, etc.

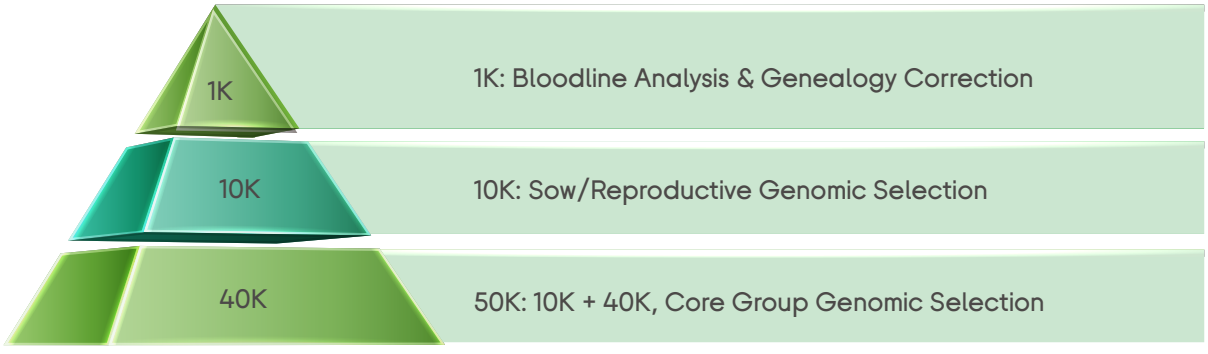


Figure 6. Applications of 50K, 40K, 10K, 1K porcine liquid chips

Results

1. There are 32,419 identical SNPs in the Porcine 50K liquid chip and the NEOGEN Porcine 50K array. 550 samples were genotyped by both method , and the detected SNPS of the two methods had a high alignment rate (Figure 7).

2. The average agreement rate of the two methods in the 550 samples is 99.05% with maximum value 99.38% and minimum value 98.42%.

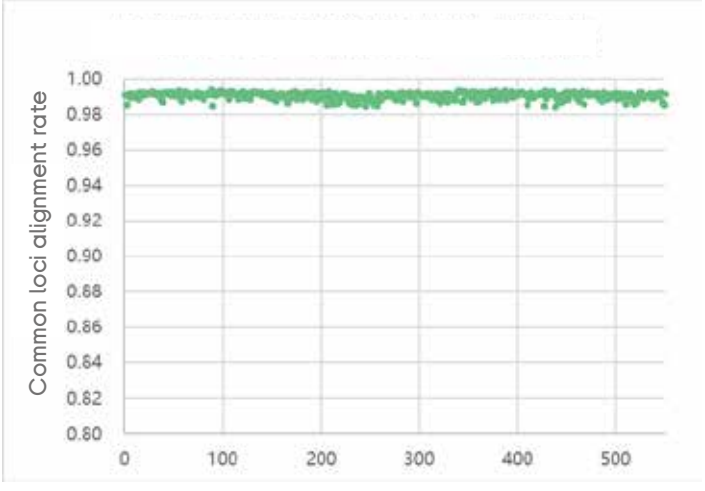


Figure 7. Alignment rate of 550 samples detected by Porcine 50K liquid chip and the NEOGEN Procine 50K array

Case 4: RAD sequencing
——Comparison of RAD sequencing and DNA arrays based on BGISEQ-500

Experimental design

453 Yorkshire pigs were subjected to genotyping with RAD-Seq (1.4 G) and GeneSeek 50K SNP array, which targeted at eight phenotypes of the pigs, including birth weight, body height, body length as well as corrected age, corrected daily gain, corrected backfat thickness, corrected eye muscle area and lean meat percentage at 100 kg.

Results

1. RAD-Seq based on BGISEQ-500 achieved average sequencing reads per individual 7.16 M (single-end), average sequencing depth 5.65×, average genome alignment rate 95%, and average coverage rate 8.3%.

Table 2. RAD-seq library sequencing data

	Raw reads	Clean reads	Map rate	Coverage	Depth
Min	472542	446967	0.87	0.024	1.40
Max	24455828	20958535	0.97	0.113	13.40
Mean	7160322	6613671	0.95	0.083	5.65

Note: Since the number of reads at both ends of the PE100 sequencing reads is the same, only the single-end reads are counted.

2. RAD-Seq generated 7.16 M data and detected 139,634 SNPs, among which, 20,294 SNPs were novel. In comparison, the 50K SNP array detected 45,180 SNPs. Besides, RAD-seq detected more SNP genotypes with low allele frequency than the SNP array.

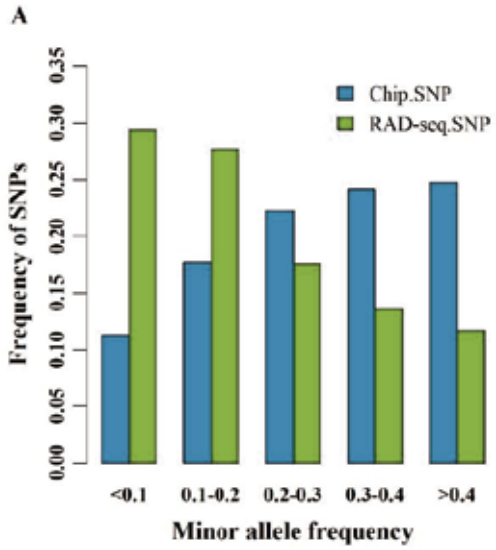


Figure 8. Minor allele frequency

3. Accuracy evaluation: The RAD-seq data was imputed with the deep resequencing data (19×) of large white pigs from the same farm, and the accuracy evaluation found that the imputation accuracy was about 0.85.

Conclusions

Both RAD-seq and SNP array technology could find genome wide SNPs. The number of SNPs identified by RAD-seq is three times that of SNP array while at a similar cost. Due to the different sequencing depths among individuals, there were differences in the number and location of SNPs obtained by RAD-seq. Some individuals have allele loss due to mutation at the restriction site. Therefore, it is recommended to use high-depth resequencing data from the same population to impute the RAD-seq data to obtain genotype data with high credibility.

Reference

Li Yong. Evaluation of various SNP genotyping methods in pig genomic selection and genome-wide association study [D]. Huazhong Agricultural University, 2020.

Contact Us

MGI Tech Co., Ltd.
Address: Building 11, Beishan Industrial Zone, Yantian District, Shenzhen, CHINA 518083
Email: MGI-service@mgi-tech.com
Website: https://en.mgi-tech.com/
Tel: 4000-966-988
Version: April 2022 | MGPD113810200-01

https://www.linkedin.com/company/mgi-bgi

https://twitter.com/MGI_Technology

*Unless otherwise informed, StandardMPS and CoolMPS sequencing reagents, and sequencers for use with such reagents are not available in Germany, USA, Spain, UK, Hong Kong, Sweden, Belgium, Italy, Finland, Czech Republic, Switzerland and Portugal.