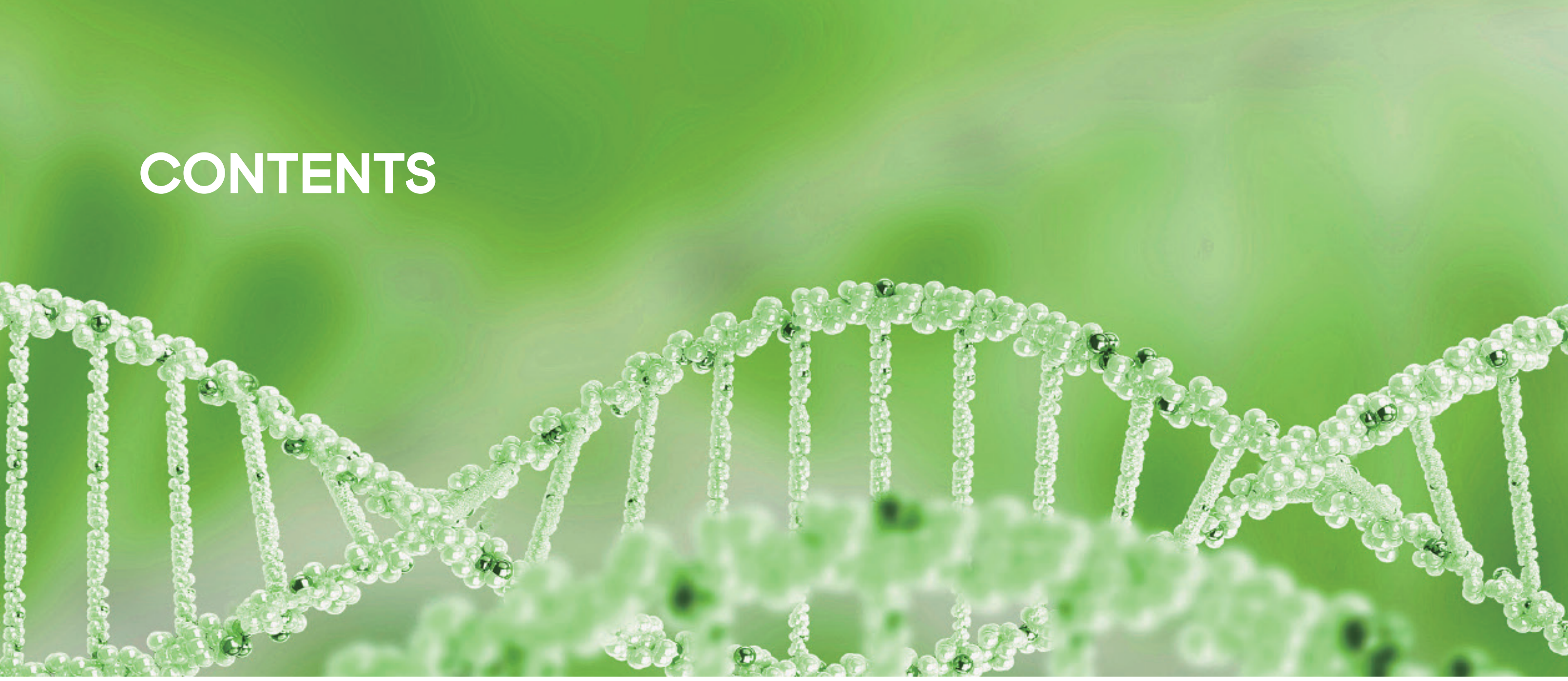
Proposal for Agrigenomics Projects

based on MGI sequencing platforms

CONTENTS



Molecular breeding scheme based on DNBSEQ™ sequencing platforms

Automated sample processing system to speed up the pre-processing efficiency of breeding samples

-Automatic Sample Transfer Processing System MGISTP-7000

-Automated Sample Preparation System MGISP-960

DNBSEQ™ sequencing technology accelerates the process of molecular breeding

-Technical principles and platforms

-Recommendations for molecular breeding and sequencing strategies

-Molecular Breeding Analysis System

Research status and development trend of agrigenomics

In recent years, scientists have made significant progress in the field of basic research on animal husbandry and molecular breeding of crops, breaking through a series of important scientific challenges and key technologies, and forming a good production-university-research model with breeding enterprises. In the field of agriculture, high-throughput sequencing-based animal and plant genomics research mainly involves whole genome, genome resequencing, RNA, small RNA, single cell and other sequencing methods. The sequenced species cover crops, vegetables, fruit, poultry, livestock, fish, model animals, insects, etc. Researchers have increased the efficiency of breeding by hundreds or even thousands of times and shorten the breeding cycle by two thirds based on high-throughput sequencing from genome mapping to gene mining and molecular breeding.

Sequencing technology and genome mapping of important agricultural species

Genomics research is a discipline that studies how genes and genetic information in a species' genome are organically combined and determine their functions. This discipline breaks through the previous model of research at the level of individual genes, and opens up a new field targeting the structure, expression and interaction of the genome. With the emergence of series of animal and plant genome research, agriculture genomics research has also entered the post-genome era. Our cognitive level has also entered the molecular structure, product and biochemical mechanism from the original phenotypic description.

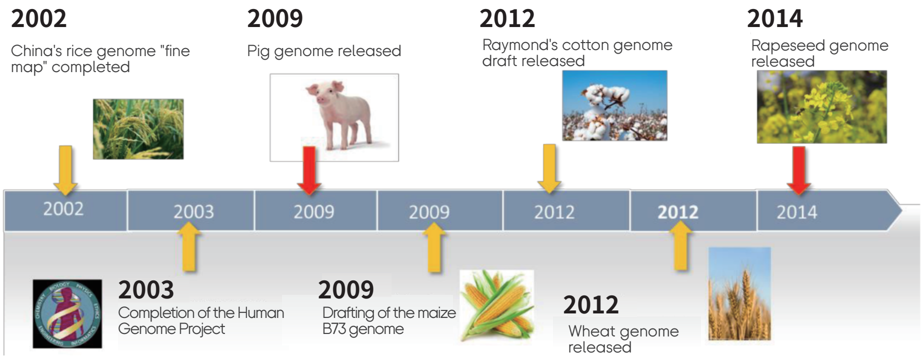

2002
China's rice genome "fine map" completed
2009
Pig genome released
2012
Raymond's cotton genome draft released
2014
Rapeseed genome released
2002
2003
2009
2009
2012
2012
2014
2003
Completion of the Human Genome Project
2009
Drafting of the maize B73 genome
2012
Wheat genome released

Figure 1. The completion of the drawing of the genome map of important crops and livestock

The completion of many animal and plant genome projects has been establishing a rich genetic information database for biodiversity, animal and plant evolution mechanisms, molecular breeding research, etc. It also provides unprecedented basic resources for global animal and plant researchers. The completion of various drawing of the species genome map marks the entry of the species breeding into the era of agricultural genomics, or agrigenomics. By sequencing a variety of important agricultural animal and plant species such as rice, pig, corn, cotton, wheat, and rape, and analyzing the reference genome map, it is helpful for the breeding improvement and ecological management of agricultural species, and is conducive to the study of the environment of agricultural species. It is also beneficial to the study of adaptation of agricultural species to environment and the molecular mechanisms behind it, which will be important for protection of endangered species.

Sequencing technology and the accelerated process of molecular breeding

According to scientific assessments of agriculture production made by the International Food and Agriculture Organization (FAO) and developed countries on aquaculture production, the contribution rate of species is 35% to 65%, with an average of 45%. With the progress of various plant genome projects, this proportion will increase substantially. Variety is the primary key to the development of animal and plant agriculture. The propagation of excellent varieties is one of the core contents of agricultural animal and plant breeding, which directly affects the efficiency of animal and plant agricultural production. With improved seeds and high-speed propagation technology, greater output can be achieved with the same input conditions.

Sequencing technology have been widely used in the construction of species genome maps. Besides, resequencing, transcriptome sequencing, comparative genome analysis and other technologies developed based on sequencing technology have also been applied to the mining of molecular markers of important agronomic traits. With continuous reduction of the cost of sequencing nowadays, sequencing technology has also begun to be widely used in molecular breeding to improve breeding efficiency and allow for faster breeding cycles.

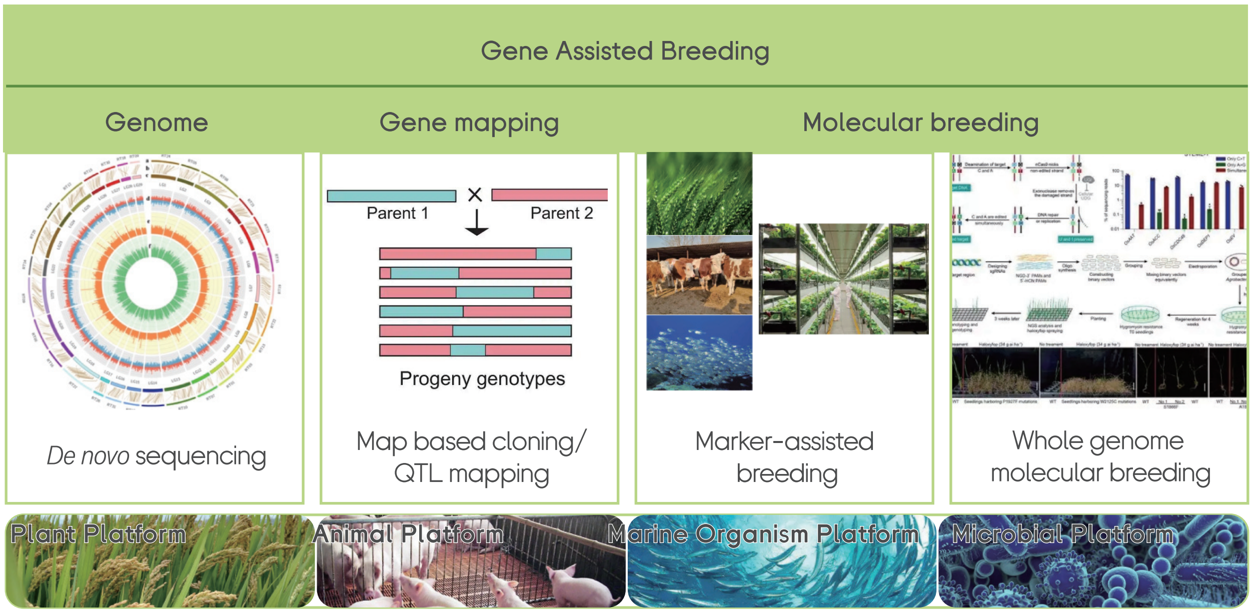


Figure 2. Sequencing technology used in various stages of gene-assisted breeding

Traditional breeding is a method of phenotypic selection. This method selects the best line according to traits. The method is simple, but it has a long breeding cycle and low accuracy. With the development of molecular biology, the identification of genes related to some important traits has boosted the rapid application of marker-assisted selection. Marker-assisted selection is based on molecular markers to assist in the selection of optimal individuals. However, most economic traits are complex quantitative traits controlled by multiple genes and environmental factors, which also greatly limits the application of marker-assisted selection.

With the advances in sequencing technology and the emergence of DNA arrays, scientists have developed a genomic selection (GS) method, which is based on genome-wide markers to assist in the selection of optimal lines or individuals for multiple traits. This method uses the phenotypes and marker genotypes in the reference population for estimation. Once an estimate of the marker effect is obtained, individual genotypes can be used to predict individual breeding values. It can assist in the selection of best breeders with multiple traits and has great advantages in improving the accuracy of breeding value estimation and shortening the generation interval. The breeding efficiency has been increased hundreds or even thousands of times, and the breeding cycle has been shortened to two-thirds of the original (Figure 3).

Basic scientific research in agriculture based on the DNBSEQ^{TM} sequencing platforms

Several sequencing platforms have been developed based on DNBSEQ^{TM} sequencing technology and the applications covers DNA sequencing (including whole-genome resequencing) of various genome sizes (such as animals, plants, microorganisms, etc.), RNA-seq (including small RNA, mRNA, transcriptome), ChIP-seq, sequencing of enriched target regions, DNA amplicon sequencing, verification of mutation sites, etc. To sum up, its applications cover various fields of scientific functional genomics research:

DNA level: a. Whole genome sequencing; b. Microbial amplicon sequencing and metagenomic sequencing; c. Large-scale screening of gene mutations and gene polymorphism analysis; d. Whole genome methylation sequencing; e. ATAC-SEQ Sequencing;

RNA level: a. Transcriptome sequencing; b. Whole transcriptome sequencing; c. Long non-coding RNA sequencing; d. Alternative RNA splicing and RNA variation, such as RNA SNP, etc.; e. Single cell transcriptome Sequencing.

Genome map construction

De novo sequencing of sepecies genome refers to the sequencing and construction of a new genome or transcriptome, which contributes to new understanding of an organism and can also be used for genome comparison studies. On the other hand, whole-genome sequencing helps to identify genomic variations such as SNPs, indels, copy number, and other sample structural variations associated with a common reference sequence. However, the accuracy of third-generation sequencing technology using long-read length sequences for *de novo* is limited while second-generation sequencers using short-read length technology have high accuracy but low genome *de novo* integrity.

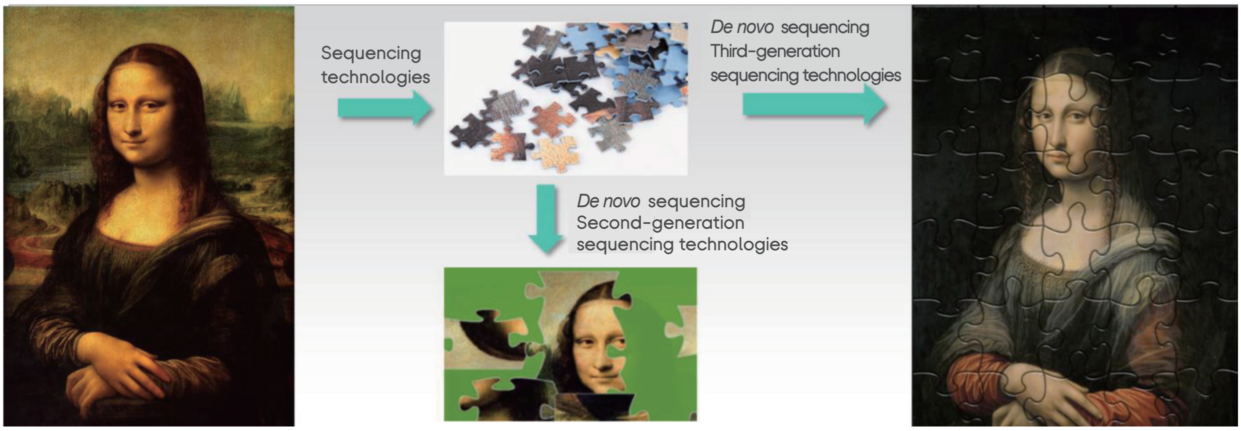


Figure 4. The application effect of second- and third-generation sequencing technologies in *de novo* sequencing

MGI single-tube long fragment sequencing (stLFR)

stLFR (single tube Long Fragment Read): single tube long fragment sequence, that is, the short-read sequencing fragments derived from the same long DNA fragment are labeled with the same molecular tag. This technology is independently developed by MGI based on the DNBSEQ™ platform. The long-fragment reading technology realizes the acquisition of long-fragment DNA information based on high-precision short-read sequencing, and the length of the read sequence can be as high as 10k~300k.

Using MGIEasy stLFR Library Prep kit combined with MGI's DNBSEQ™ , the world's leading sequencing technology, stLFR enables high quality small variants calling, phasing diploid genomes, detection of structure variations and other long read applications.

With a small amount of input HMW genomic DNA (1-1.5 ng) added to a single tube, the gDNA molecules will be barcoded with 30 million barcoded beads. stLFR technology can specifically co-barcode more than 8 million long fragments ranging from 20,000 to 300,000 bp in a single tube (Figure 5). By using the stLFR assembly software to assemble the huge barcode information efficiently, a more perfect genome assembly effect can be obtained, making the assembly of animal and plant genomes simpler, faster, and more economical.

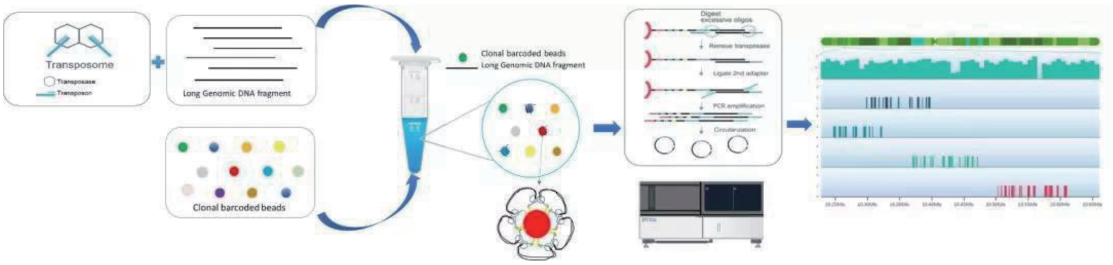


Figure5. stLFR

stLFR case

Comparison of three long read methods for sequencing and assembly of the macadamia nut genome[1]

In the article "Comparison of long read methods for sequencing and assembly of a plant genome" published by the University of Queensland, Australia, three long-read sequencing technologies were used to perform the *de novo* assembly of a plant genome, Macadamia jansenii, including Pacific Biosciences (Sequel I), Oxford Nanopore (PromethION, MinION) and MGI (stLFR + MGISEQ-2000 (DNBSEQ-G400)). The stLFR assembly contained the fewest mismatches and indels and highest accuracy was observed; while the ONT assembly results were significantly different before and after correction (Figure 6).

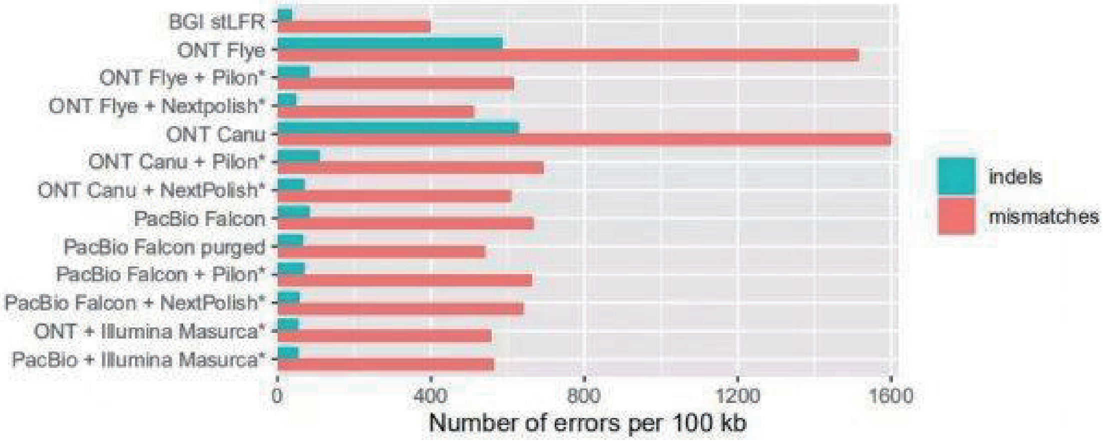


Figure6. Comparison of base accuracy of assembly results

There was no significant difference in the genome completeness of the three methods: ONT+Illumina, PacBio and stLFR+ONT, all above 92% (Figure 7).

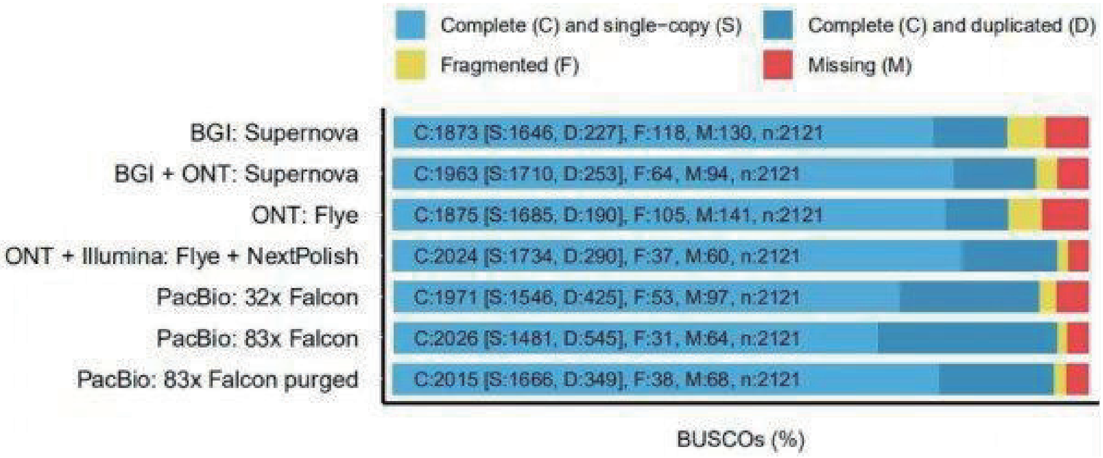


Figure 7. Assembly Completeness Assessment

From the comparative analysis, all the 3 long-read sequencing technologies, Pacific Biosciences (Sequel I), Oxford Nanopore (PromethION) and MGI stLFR, can achieve very good assembly results. Relatively speaking, MGI stLFR required the lowest amount of DNA input which was only at nanogram level, while the assembly contained the fewest mismatches and InDels and was most accurate. It will be especially suitable for precious samples or extreme low amount samples, yet with significant advantage on sequencing cost.

Development of genome-wide molecular markers

Using whole-genome resequencing and RNA sequencing, researchers can quickly conduct resource census screening to look for a large number of genetic differences, and perform genetic evolution analysis and predict candidate genes for important traits.

Whole-genome resequencing

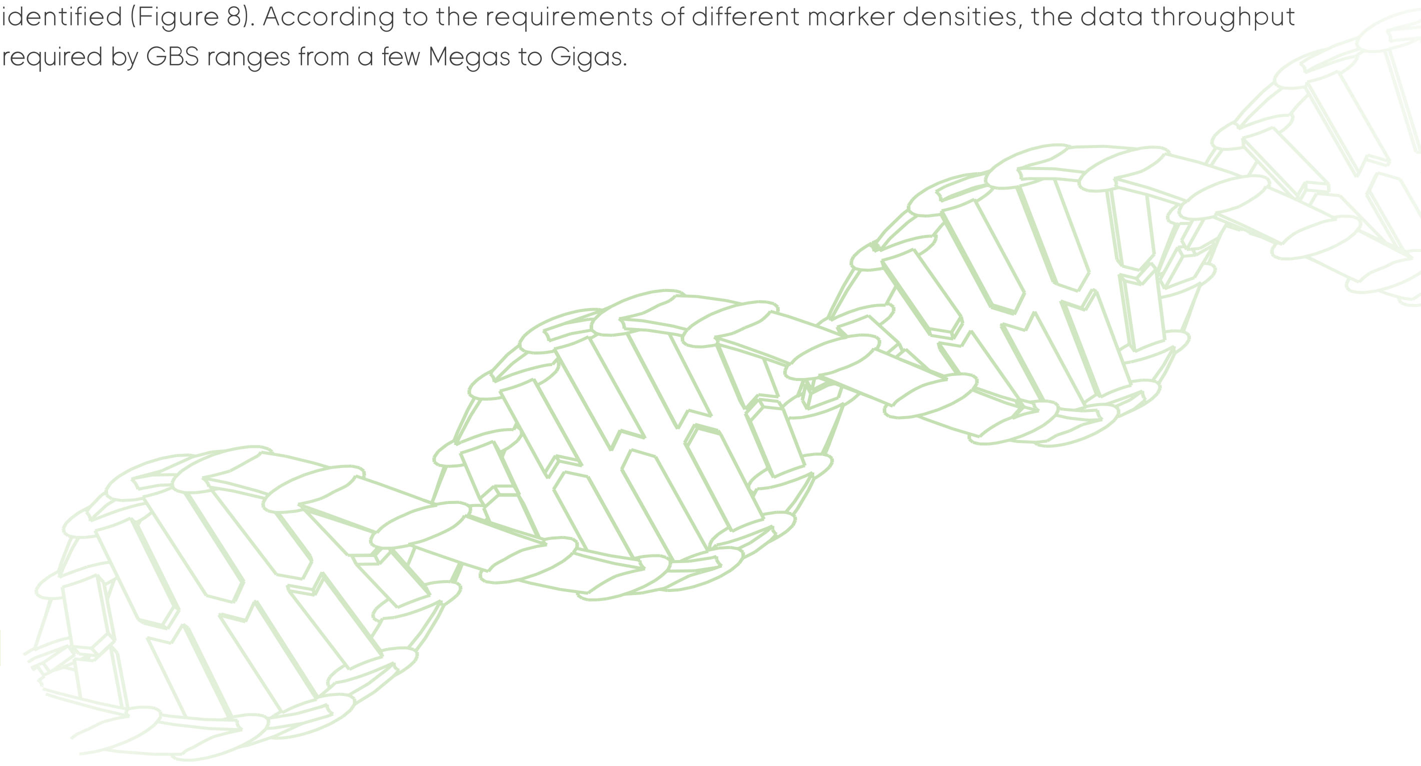
Genome resequencing of core germplasm collection: Core collection is a core subset of germplasm resources, which preserve the genetic diversity of the entire population to the greatest extent, while represent the geographic distribution of the entire population. Resequencing of core collection can provide an in-depth understanding of the genetic diversity of genes and genotypes on the largest scale. Besides, it has important academic and practical significance for promoting germplasm exchange, utilization and management.

Population genome resequencing: for the known species genomes, the genome resequencing of different individuals can discover differences among individuals within a population at the genomic level. A large number of SNPs, InDels, structural variations (SVs) and other variation information can be found with this method, and the genetic characteristics of the biological population can be obtained. It is instrumental to study the evolutionary history and environmental adaptability of species at population level. Whole-genome resequencing can also help quickly discover genetic variation related to important traits of animals and plants, and shorten the experimental period of molecular breeding.

Simplified Genome Sequencing

Simplified genome sequencing is currently the most efficient and practical means for population analysis. RAD (Restriction Associated DNA Sequencing, RAD-seq) and GBS (Genotyping by sequencing) are common methods for simplified genome sequencing.

GBS in the narrow sense is based on enzyme digestion, that is, by sequencing the tags after restriction endonuclease digestion, SNPs with high density and uniform distribution in the genome can be rapidly identified (Figure 8). According to the requirements of different marker densities, the data throughput required by GBS ranges from a few Megas to Gigas.



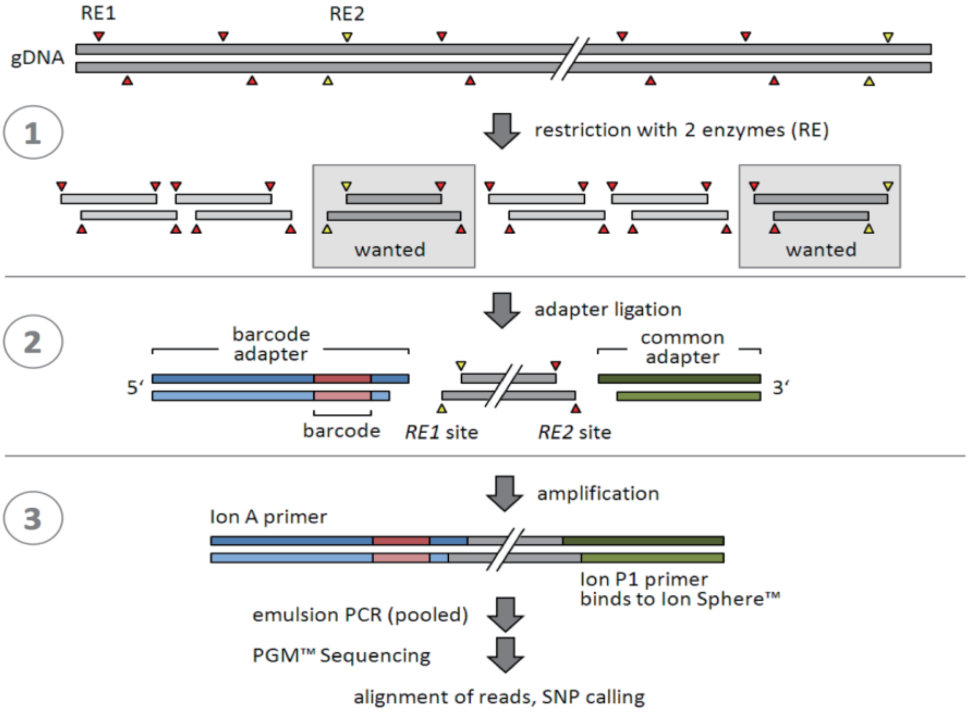


Figure 8. Enzyme digestion based GBS experimental workflow

RNA sequencing

For species with reference genome, RNA-seq results can be compared with the DNA sequence data to obtain analysis results such as gene expression, alternative splicing, gene structure optimization, and new gene discovery.

For species without a reference genome, *de novo* transcriptome sequencing studies can be performed, where *de novo* assembly of the obtained sequencing reads was conducted and used to obtain the unigenes data of the species. This can be effectively used for the discovery of new genes and the development of new molecular markers for numerous and diverse germplasm resources, especially species with large and complex genomes.

Application cases

With the collaboration of Shenzhen BGI Life Sciences Research Institute, Netherlands Genetic Resources Center, Shenzhen National Gene Bank, Huazhong Agricultural University and other institutes, a research paper titled "Whole-genome resequencing of 445 Lactuca accessions reveals the domestication history of cultivated lettuce" was published in Nature Genetics. They carried out the whole-genome resequencing of 445 Lactuca accessions from 47 countries around the world, including major lettuce crop types and wild relative species, and generated a comprehensive map of lettuce genome variations.

Through phylogenetic analysis, the research team found that all cultivated lettuce originated from an independent domestication event. The principal component analysis and population structure analysis showed that the cultivated lettuce had the closest genetic distance to the wild lettuce population in the Caucasus and the Mesopotamia. From this, it is inferred that cultivated lettuce is very likely to originate in the Caucasus and the Mesopotamia. Through effective population size analysis, it was found that both cultivated lettuce and wild lettuce experienced population decline 10,000 years ago, which may have been caused by drastic changes in the environment. From 4000 BC, the effective population size of cultivated lettuce showed a more dramatic decline, suggesting that lettuce was undergoing domestication. After being domesticated by humans, lettuce first spread to ancient Egypt and gradually evolved into today's oil lettuce. It was introduced to southern Europe in ancient Roman times, and after crossing with local wild lettuce, it began to be grown and eaten as leaf lettuce.

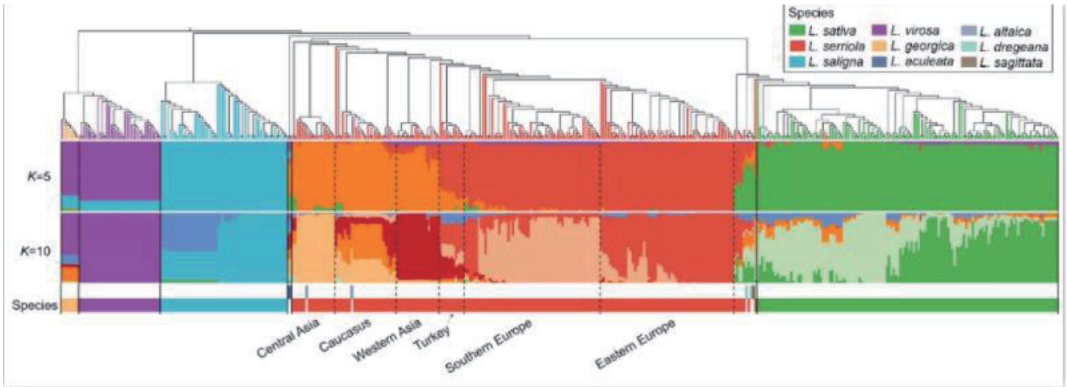


Figure 9. Population analysis of cultivated lettuce (shown in green) and wild relatives

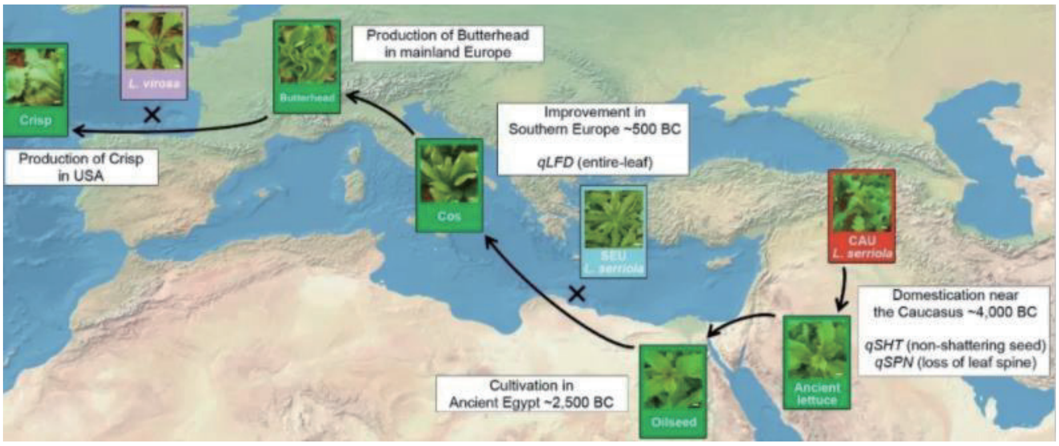
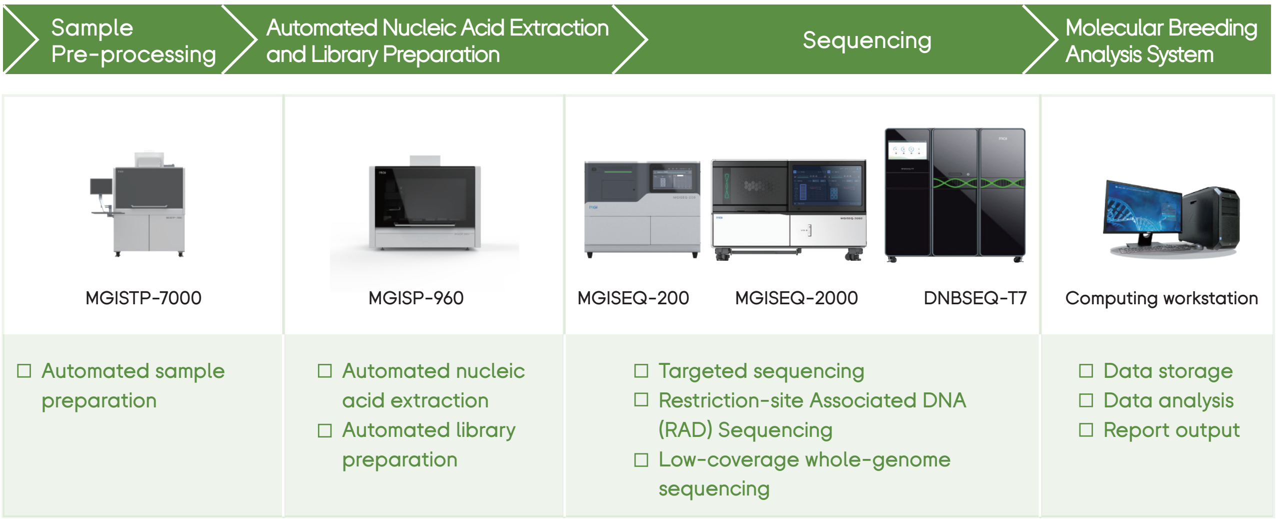

Production of Butterhead in mainland Europe
L. virosa
Crisp
Butterhead
Production of Crisp in USA
Improvement in Southern Europe ~500 BC
qLFD (entire-leaf)
Cos
SEU L. serriola
CAU L. serriola
Domestication near the Caucasus ~4,000 BC
qSHT (non-shattering seed)
qSPN (loss of leaf spine)
Ancient lettuce
Cultivation in Ancient Egypt ~2,500 BC
Oilseed

Figure 10. The history of lettuce domestication

Cultivated lettuce was first domesticated from the wild lettuce (red box in the picture) in Caucasus and Mesopotamia, and then it was introduced to ancient Egypt and cultivated for oil production, and then the oldest leaf lettuce appeared in southern Europe—the shallow-rooted lettuce, also known as romaine lettuce, because it was widely grown in ancient Rome.

The research project was conducted on the BGISEQ-500 sequencing platform, which realizes the introduction of high-quality germplasm resources, the digitization of resources, and the mining of important functional genes.

Molecular breeding scheme based on DNBSEQ™ sequencing platforms



Sample Pre-processing

Automated Nucleic Acid Extraction and Library Preparation

Sequencing

Molecular Breeding Analysis System

MGISTP-7000

MGISP-960

MGISEQ-200

MGISEQ-2000

DNBSEQ-T7

Computing workstation

- □ Automated sample preparation

- □ Automated nucleic acid extraction
- □ Automated library preparation

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

- □ Data storage
- □ Data analysis
- □ Report output

Figure 11. Molecular breeding scheme

Automated sample processing system to speed up the pre-processing efficiency of breeding samples

Automatic Sample Transfer Processing System MGISTP-7000

Molecular breeding usually need a lot of manpower to process sample transfer in the early stage of a project with a large sample size, and the manual operations are slow and prone to errors. Automated sample transfer system MGISTP-7000 can be used to process blood samples from livestock, poultry and aquatic products efficiently.

MGISTP-7000 is a easy-to-use high-throughput automated sample transfer processing system. It integrates tube decapping, tube recapping, barcode identification, automated liquid transfer, and negative pressure protection. It can transfer 192 samples from airtight sample tubes or plain tubes to 96-Well Microplates in 40 min. Compared with manual operation, the MGISTP-7000 can increase the efficiency by 3-4 times. A large size of samples can be transferred easily even in the event of short of manpower.

Automated Sample Preparation System MGISP-960

MGISP-960 High-throughput Automated Sample Preparation System is a flexible, fully automated workstation with 96-channel pipette and can be used in various applications. In the NGS field, most nucleic acid extraction and library preparation processes can be automated. MGISP-960 can be adopted to process samples in batches, eliminating the need for repeated manual operations, improving the stability of NGS library preparation, reducing the total cost, and comprehensively improving the overall work efficiency of the laboratory.

MGISP-960 can be used for molecular breeding application together 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.

DNBSEQ™ sequencing technology accelerates the process of molecular breeding

Technical principles and platforms

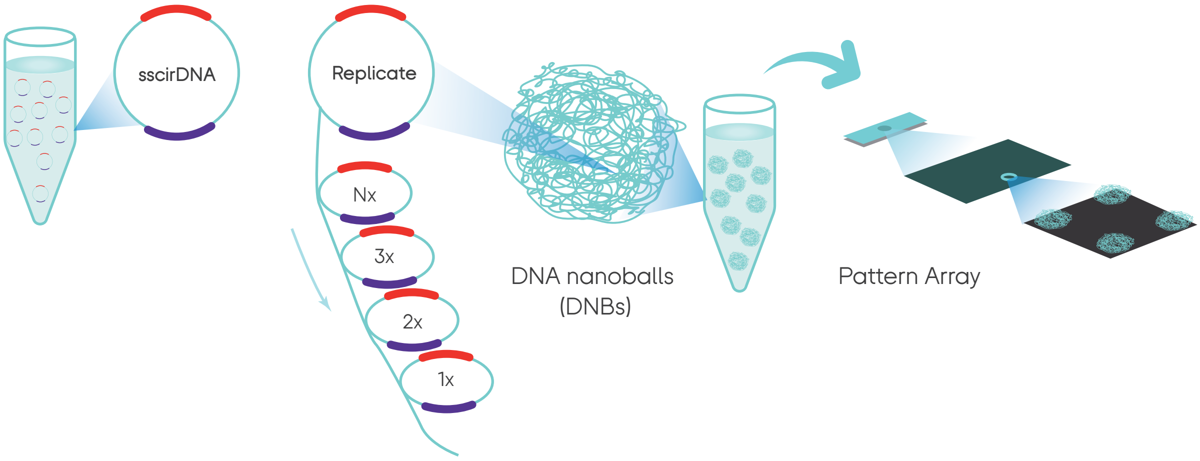


Figure 12. Rationale of DNBSEQ™ sequencing technologies

Increased accuracy	Decreased duplicates	Reduced index hopping
> 99.9% SNP Precision/Sensitivity > 99% Indel Precision/Sensitivity	<3%	0~0.0004%
* MGI PCR-Free library for detection accuracy and sensitivity of NA12878 InDel		

The core technology of DNBSEQ™ sequencing technologies is the use of cPAS sequencing technology to establish a direct link between chemical and numerical information. After DNA nanoballs (DNBs) are prepared from DNA fragments, DNBs are immobilized on a patterned array flow cell, and fluorescent labeled dNTP probes are subsequently incorporated. Bases are called in real time with the fluorescence signal excited by the binding of each base. Compared with other sequencing technologies, the use of this technology has high accuracy, decreased duplicates and extreme low index hopping rate.

MGI developed several sequencers with different throughputs to meet different market demands based on DNBSEQ™ sequencing technology, such as:

DNBSEQ-G50 is a compact and flexible benchtop genetic sequencer and creates a perfect balance between speed and affordability. It is suitable for disease prevention, clinical research, and hospital medical diagnostic laboratories that have data volume requirements.

DNBSEQ-G400* is a versatile high-throughput benchtop sequencer providing users with comprehensive, flexible and efficient sequencing options. It supports a wide range of applications including scientific research, clinical research, disease prevention, environment studies and agriculture, etc.

DNBSEQ-T7* is an ultra-high-throughput sequencer and can generate 6T of high quality data in 28 hours when 4 flow cells are fully loaded. It is widely used in scientific and research services, medical services and breeding services.

Product Model	DNBSEQ-T7	DNBSEQ-G400	DNBSEQ-G50
Features	Ultra-high Throughput	Adaptive	Effective
Applications	Whole Genome Sequencing, Deep Exome Sequencing, Transcriptome Sequencing, and Targeted Panel Projects.	WGS, WES, Transcriptome sequencing, etc.	Small whole genome sequencing, targeted DNA/RNA panels, low-pass whole genome sequencing
Flow Cell Type	FC	FCL & FCS	FCL & FCS
Lane / Flow Cell++	1 lane	2 or 4 lanes	1 lane
Operation Mode	Ultra-high Throughput	High Throughput	Medium Throughput
Max. Throughput / RUN	6 Tb	1400 Gb	150 Gb
Effective Reads / Flow Cell	5000 M	1500-1800 M	500 M/100 M
Average run time	24-30 hours for PE150 sequencing	PCS: 13-37 hours; FCL: 14-109 hours	9-40 hours
Max. Read Length	PE150	SE400/PE200	PE150
Min. Read Length	PE100	SE50	SE50

Figure 13.   Performance parameters of different sequencers based on DNBSEQ™ sequencing technology

• Targeted sequencing

-Multiplex PCR (Customized ATOPlex platform)

Multiplex PCR is a new type of PCR technology developed based on conventional PCR and can detect multiple targets simultaneously to obtain the sequence with multiple target genes in the sample, which not only saves time and cost, but also can obtain more information from extreme low samples. It is widely used in many applications such as pathogens, tumors, agriculture and husbandry.

ATOPlex platform, based on MGI's self-developed ultra-high multiplex PCR technology, offers targeted library prep kits customized service with technologies such as Auto-workflow, Trace-samples, One-tube and Pure-PCR. It covers DNA, RNA and DNA methylation and can be applied in medicine, research, forensic, agriculture, DTC (direct to customer) and other gene detection demand (Figure 14).

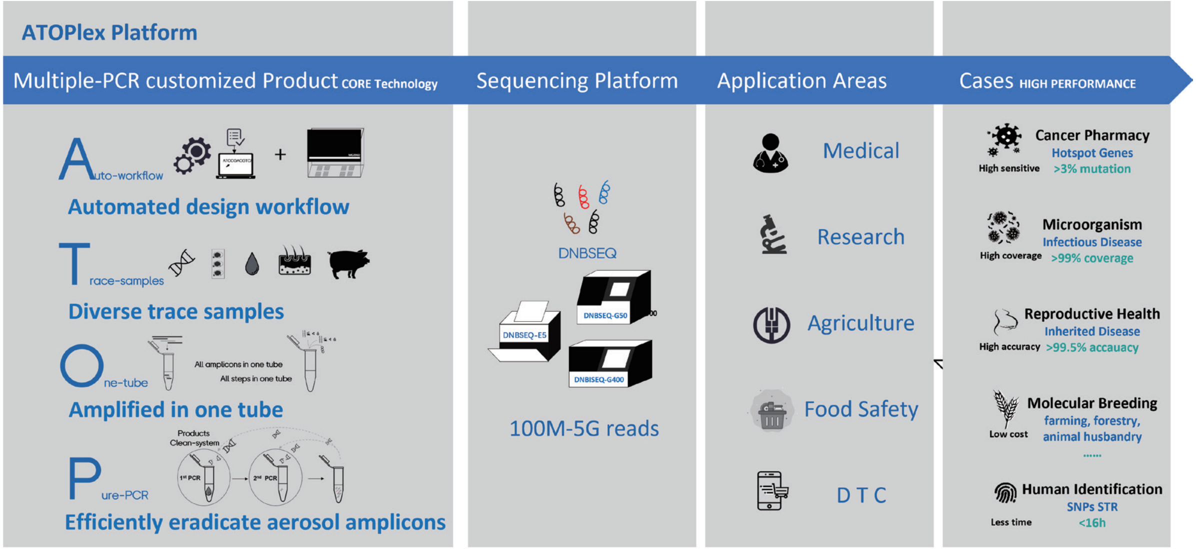


Figure 14. Customized ATOPlex Platform

ATOPlex - Scalable Genotyping Technology

ATOPlex can design customized agricultural genome detection products. Each sample can detect 100-5000 regions, and at least 100-5000 SNPs can be detected at a time. The average depth is 100X. It supports mixed sequencing of up to 48*96 samples and is compatible with various MGI sequencing platforms.

ATOPlex - Reproducible Results

High locus call rates and greater reproducibility across a wide range of samples help ensure precise selection. ATOPlex technology has a locus call rate of over 95%, while guaranteeing >99% inter- and intra-assay reproducibility. Sequencing data can maintain high concordance (>99%) with orthogonal genotyping techniques such as microarray technology.

ATOPlex - Quick and Easy Workflow

The experiment only contains a 2-step PCR, and the library construction can be completed within 4 hours. It can be used with the MGISP-960 automated sample preparing system to further simplify the experimental process.

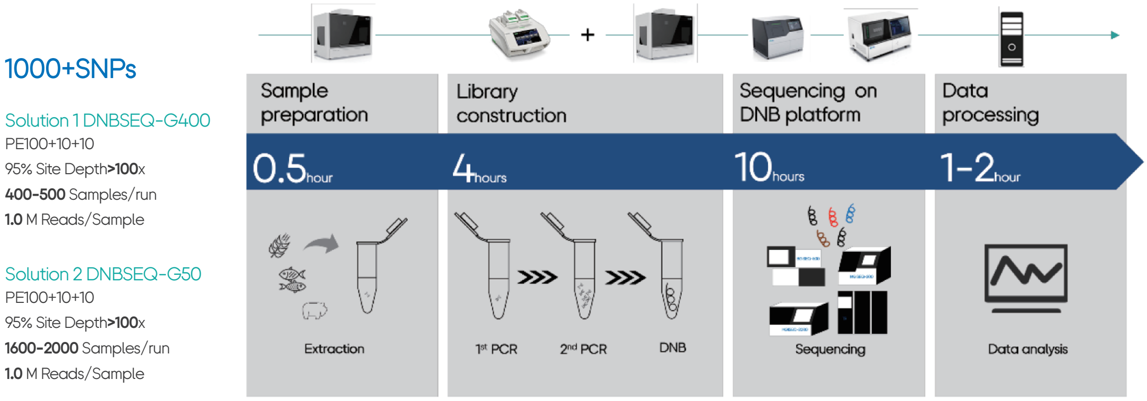


Figure 15. Customized ATOPlex Platform Workflow

Case Study: Growth Weight and Ammonia Tolerance Prediction

Panel Design for Grouper:

Product Name	Grouper Breeding Customized Panel
Amplicons	740+ plex
Amplicon size	100-200 bp
Number of detection loci	741 SNPs
Detection principle	multiplex PCR + high-throughput sequencing
Applicable sample type	gDNA
Recommended DNA input	1~10 ng DNA
Recommended reads	1.0 M reads/sample
Reading length	PE100
Compatible sequencer	DNBSEQ-T7, DNBSEQ-G400, DNBSEQ-G50
Compatible Automation system	MGISP-960RS

Conclusion:

The customized panel contains 741 independent SNPs. The prediction accuracy of grouper body weight is more than 80% and the accuracy of ammonia tolerance is more than 95%. The customized panel has more than 94% detection rate for different grouper species.

Performance data:

(1) The panel has good basic performance including >99% mapped rate, >98% on-target rate and >95% uniformity (Figure 16).

(2) The sequencing depth of all loci exceeds 500X, which can effectively detect polyploid genotypes (Figure 17).

(3) There are consistent sequencing depths in regions with different GC content (Figure 18).

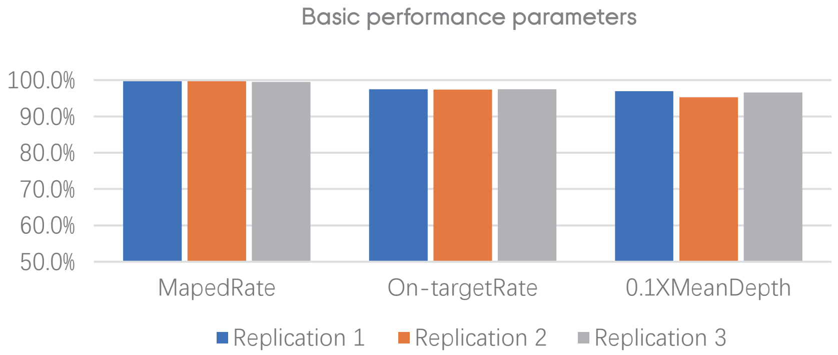


Figure 16. Basic performance parameters of customized ATOPlex panel

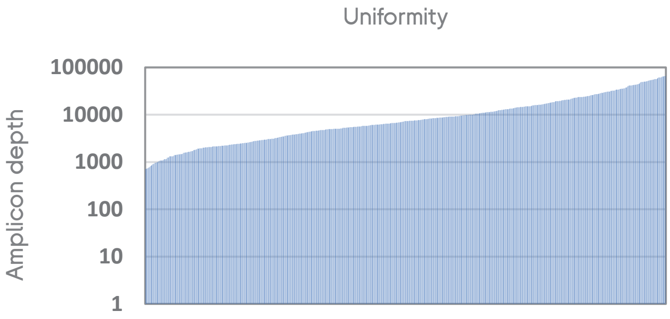


Figure 17. The uniformity of customized ATOPlex panel

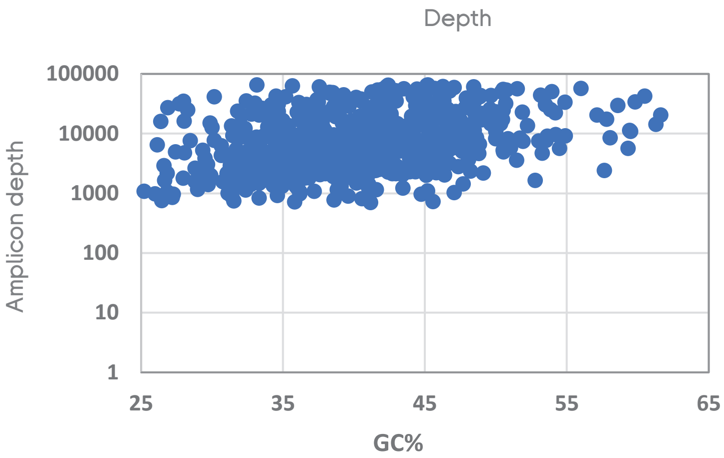


Figure 18. Depth of different GC content of customized ATOPlex panel

-Liquid-based targeted region capture (liquid chip)

The liquid-based chip is used to amplify, capture and detect target fragments under liquid conditions. It captures SNPs by using probes that are complementary to target sequences, followed by washing, amplification, library construction and sequencing, and finally obtains the genotype of the target SNPs. Its detection density and throughput are equivalent to the high-density solid-based DNA arrays.

Liquid chip technology makes marker panel design more flexible and upgradable. There is no restriction on initial sample size and marker numbers when customizing the marker panel for liquid chip. Besides, sequencing and marker genotype detection can be completed in the same tube. There is no sample size limit for a single test. New primers can be added to the system at any time or existing primers can be adjusted per required. A panel with a different marker size can be achieved based on the same high-density marker panel by adjusting the sequencing depth. This kind of flexibility is not possible for solid-based DNA arrays, as, once designed, the marker number and sample size per array are fixed.

Table 1. Comparison of solid-based DNA arrays and liquid-based chips

	DNA arrays	Liquid chip
Year of invention	2000s	2010s
Accuracy	High	Consistent with DNA arrays
Applicable number of detection markers	Low, medium and high density	Unlimited
Marking flexibility	New customization needed when adding new loci	Easily upgradable and new loci can be added at any time
Sample analysis flexibility	Gather whole plate samples for detection (e.g., 384 samples)	Flexible sequencing mode, a small number of samples can be detected, no need to gather samples for analysis
Testing equipment	Closed platform and only special testing equipment can be used	Open platform, sequencing after library construction, no special equipment requirement

Case Study: Dairy Cow 85K Liquid Chip

Huazhi Biotech, China Agricultural University and Dairy Cattle Center of Beijing Capital Agribusiness and Food Group jointly developed the 85K dairy cow breeding liquid-based chip. The liquid chip was developed by integrating Illumina and Gene-Seek arrays and the quantitative traits that are associated with milk production, milk fat content, paratuberculosis susceptibility/resistance, etc. (Table 2).

Loci source
50K chip (Illumina BovineSNP50_v3)
90K chip (Illumina Bovine90K_T_Public)
150K chip (GeneSeek GGP Bovine 150K)
Associated loci such as achondroplasia, bovine double myopathy, β -casein, k-casein and other genetic defects
Loci associated with milk production traits, milk fatty acid content, paratuberculosis susceptibility/resistance and other traits

Table 2. Loci design of dairy cow 85K liquid chip

Simplified genome sequencing

In simplified genome sequencing, only a small part of the genome is sequenced, thus the cost of sequencing the complexity of sequencing data and the time required for bioinformatics analysis all can be reduced greatly (Figure 19). With the same cost as the DNA arrays, hundreds of thousands of SNP loci can be obtained at one time.

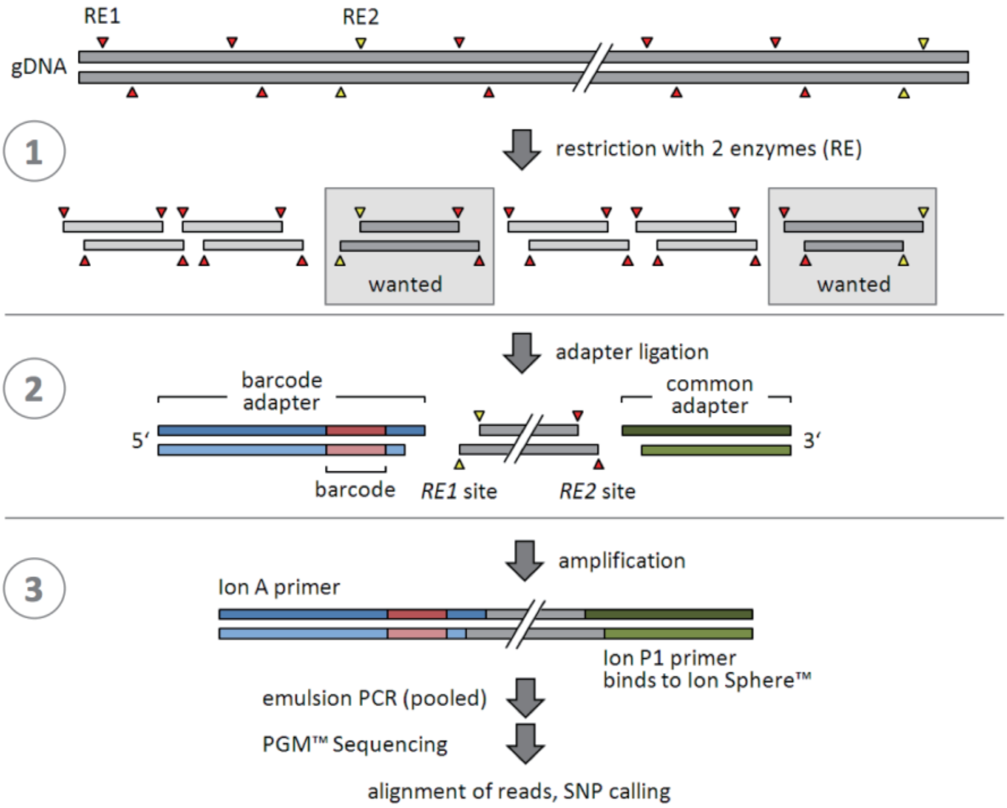


Figure 19. The experimental flow of GBS based on enzyme digestion

Case sharing: Comparison of genotype detection results based on BGISEQ-500 Yorkshire pigs simplified genome sequencing and microarray

453 Yorkshire pigs were subjected to genotyping with RAD-Seq (1.4G) 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.

(1) The average of sequencing reads per individual was 7.16M (single-end), the average sequencing depth was 5.65×, the average genome alignment rate was 95%, and the average coverage rate was 8.3% (Table 3).

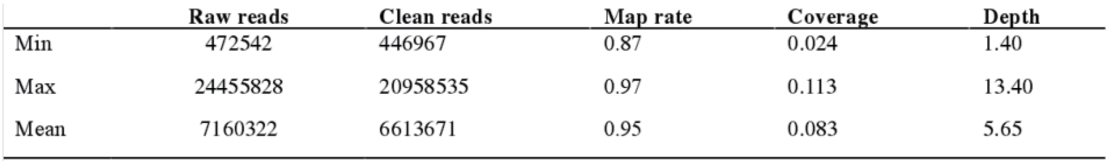
(2) RAD-Seq (7.16M) detected 139,634 SNPs, while the 50K SNP array detected 45,180 SNPs. A total of 20,294 new SNPs were found.

(3) Compared with SNP array, RAD-seq has more low-frequency SNP genotypes.

(4) Accuracy: RAD-seq data were imputed using deep resequencing data (19×) of Yorkshire pigs from the same farm and the imputation accuracy was found around 0.85.

(5) Cost: The cost of the RAD-Seq 1.4G sequencing volume is comparable to that of the 50K SNP array, but the number of SNPs identified is 3 times that of the SNP array.

Table 3. RAD library data statistics



	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 PE100 sequencing reads are the same, only the number of single-end reads is counted

Low coverage whole genome sequencing (lcWGS)

Low coverage whole genome sequencing (lcWGS) uses shotgun sequencing of genomes with 0.5-1x coverage, 'filling in' the complete sequence by computational analysis methods. While meeting the research needs, lcWGS reduces the sequencing cost by reducing the sequencing depth, and meanwhile it can obtain relatively most complete genome information when combined with the algorithm of genotype imputation. Although the sequencing depth of lcWGS is low (generally <1x), more data can be obtained , more SNPs can be found, and new SNP markers can be mined when compared with DNA arrays.

lcWGS of a large cohort has been theoretically proved to be able to obtain high-density SNP markers in the whole genome at a very low cost, thereby increasing the accuracy of QTL mapping and better mining the genetic mechanism of various diseases. It is widely used in genotype detection in molecular breeding.

Case sharing: Low-depth whole-genome sequencing of large cohorts facilitates rapid and efficient breeding of livestock [2].

The research groups from China Agricultural University and South China Agricultural University have developed a complete low-depth sequencing analysis process for livestock based on the MGISEQ-2000 (DNBSEQ-G400) sequencing platform, and analyzed the genetic structure of important economic traits in pigs. The research results titled " Accelerated deciphering of the genetic architecture of agricultural economic traits in pigs using a low-coverage whole-genome sequencing strategy" was published in the international journal Gigascience.

In the study, 2869 Duroc boars from the same breeding farm were sampled and the genome library was constructed using the self-optimized Tn5-based protocol. Low-coverage whole genome sequencing was performed on the MGISEQ-2000 (DNBSEQ-G400) platform at a mean depth of 0.73×. The Base-Var-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 20) .

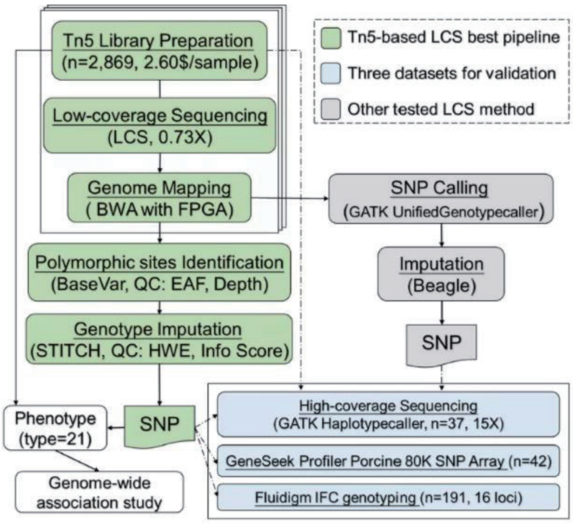

Tn5 Library Preparation
(n=2,869, 2.60$/sample)
Low-coverage Sequencing
(LCS, 0.73X)
Genome Mapping
(BWA with FPGA)
Polymorphic sites Identification
(BaseVar, QC: EAF, Depth)
Genotype Imputation
(STITCH, QC: HWE, Info Score)
Phenotype
(type=21)
SNP
Genome-wide
association study
Tn5-based LCS best pipeline
Three datasets for validation
Other tested LCS method
SNP Calling
(GATK UnifiedGenotypecaller)
Imputation
(Beagle)
SNP
High-coverage Sequencing
(GATK Haplotypecaller, n=37, 15X)
GeneSeek Profiler Porcine 80K SNP Array (n=42)
Fluidigm IFC genotyping (n=191, 16 loci)

Figure20. Low-coverage sequencing study design

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 21A). Moreover, the BaseVar-STITCH results showed even higher GC concordance and R2 values compared with the SNP chip GGP-80 data (R^2 = 0.997 and GC = 0.990). 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.

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 21 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.

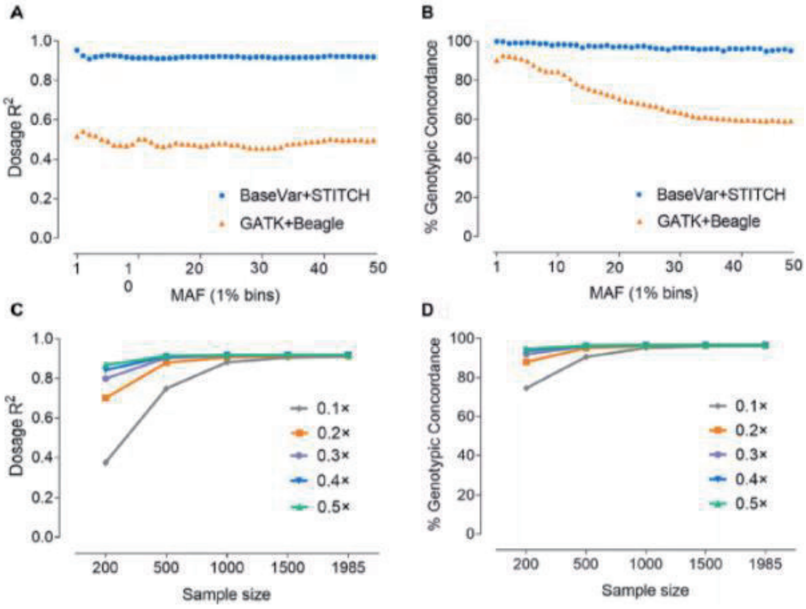


Figure 21. Performance of BaseVar-STITCH on different minor allele frequencies (MAFs) and sample sizes

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

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.

Molecular Breeding Analysis System

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

References

[1] Murigneux V, Rai S K, Furtado A, et al. Comparison of long-read methods for sequencing and assembly of a plant genome[J]. GigaScience, 2020, 9(12): giaa146.

[2] Wei T, Van Treuren R, Liu X, et al. Whole-genome resequencing of 445 Lactuca accessions reveals the domestication history of cultivated lettuce[J]. Nature Genetics, 2021, 53(5): 752-760.

[3] Li Yong. Effect evaluation and genome-wide association study of different SNP typing techniques in swine genome selection [D]. Huazhong Agricultural University, 2020.

[4] 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.

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: May 2022 | MGPV0013002

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.