华大基因 BGI | Premier Scientific Partner

Multi-level Sequencing Technologies and Applications

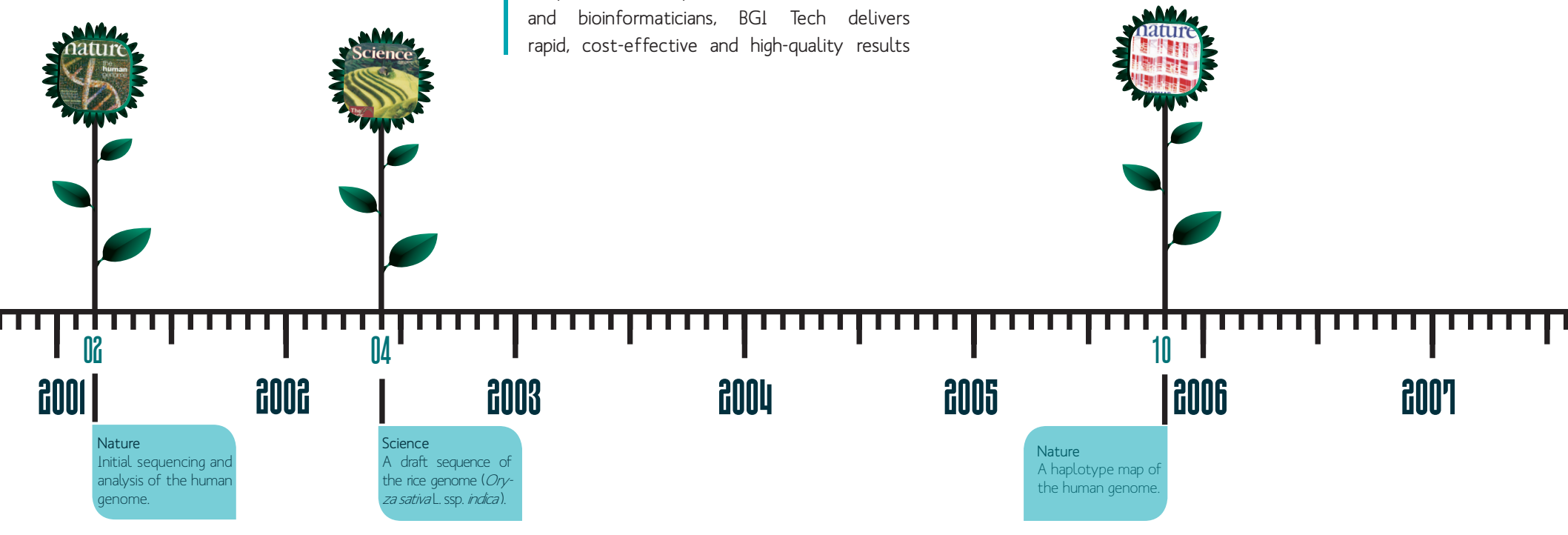
in Plant & Animal Research

华大科技 BGI tech

About BGI

BGI Tech provides advanced multi-omics and bioinformatics service solutions for its global customers in biomedical, agricultural, and environmental areas. BGI Tech is a subsidiary company of BGI, one of the largest genomics organizations in the world with significant contributions to genetic research and over 250 publications in top-tier journals such as Nature and Science. Equipped with the industry's broadest array of cutting-edge technologies coupled with an experienced team of scientists and bioinformaticians, BGI Tech delivers rapid, cost-effective and high-quality results that enable researchers to achieve scientific breakthroughs.

These successful experiences enable BGI Tech to decode the whole genomes of various plants and animals. To date, BGI Tech has sequenced 656 Plant & Animal reference genomes, including 421 animal genomes and 235 plant genomes.



At BGI, we provide high quality services that leverage the industry's best technology, economies of scale, and expert bioinformatics resources. This enables our customers and collaborators to quickly migrate from samples to discovery.

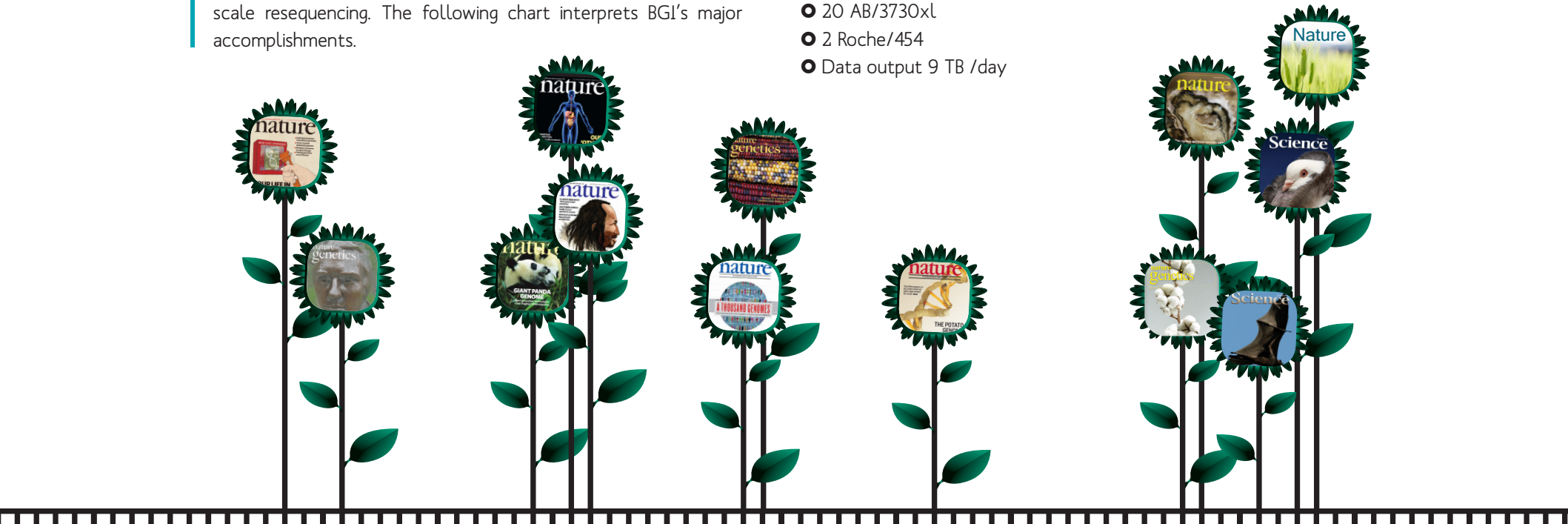
We are highly experienced in *de novo* sequencing and large-scale resequencing. The following chart interprets BGI's major accomplishments.

Sequencing platform*

- 132 Illumina/HiSeq 2000
- 5 Illumina/HiSeq 2500
- 2 Illumina/MiSeq
- 37 Life Tech/Ion Torrent
- 20 AB/3730xl
- 2 Roche/454
- Data output 9 TB /day

Computing platform*

- 212 TB FLOPS
- 37.2 TB Memory
- 20 PB Storage



Nature
The diploid genome sequence of an Asian individual.

Nature Genetics
The genome of the cucumber, *Cucumis sativus* L.

Nature
The sequence and *de novo* assembly of the giant panda genome.

Nature
A human gut microbial gene catalogue established by metagenomic sequencing.

Nature
Ancient human genome sequence of an extinct Palaeo-Eskimo.

Nature
A map of human genome variation from population-scale sequencing.

Nature Genetics
Genome-wide patterns of genetic variation among elite maize inbred lines

Nature
Genome sequence and analysis of the tuber crop potato

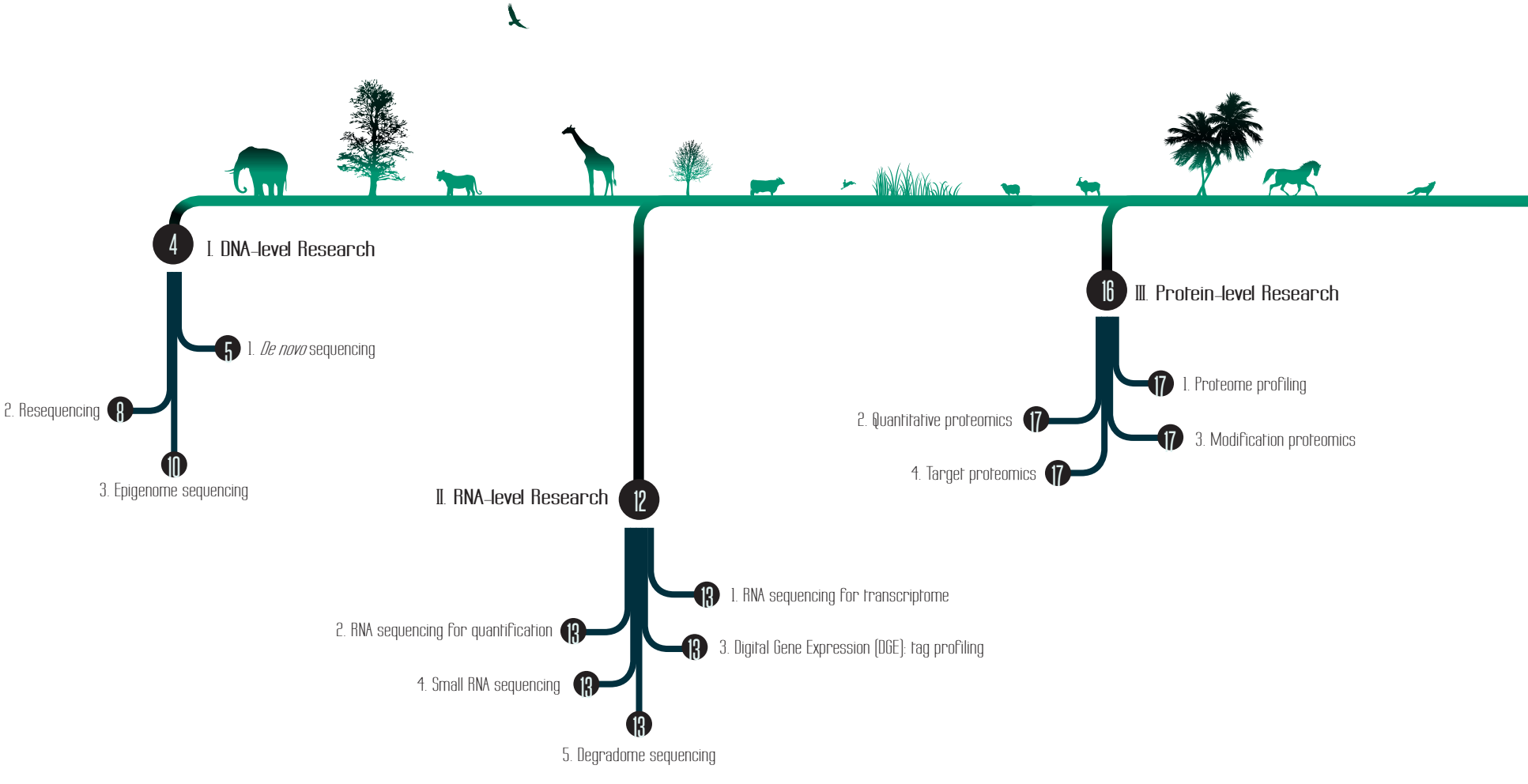
Nature
The oyster genome reveals stree adaptation and complexity of shell formation

Nature Genetics
The draft genome of a diploid cotton Gossypium raimondii

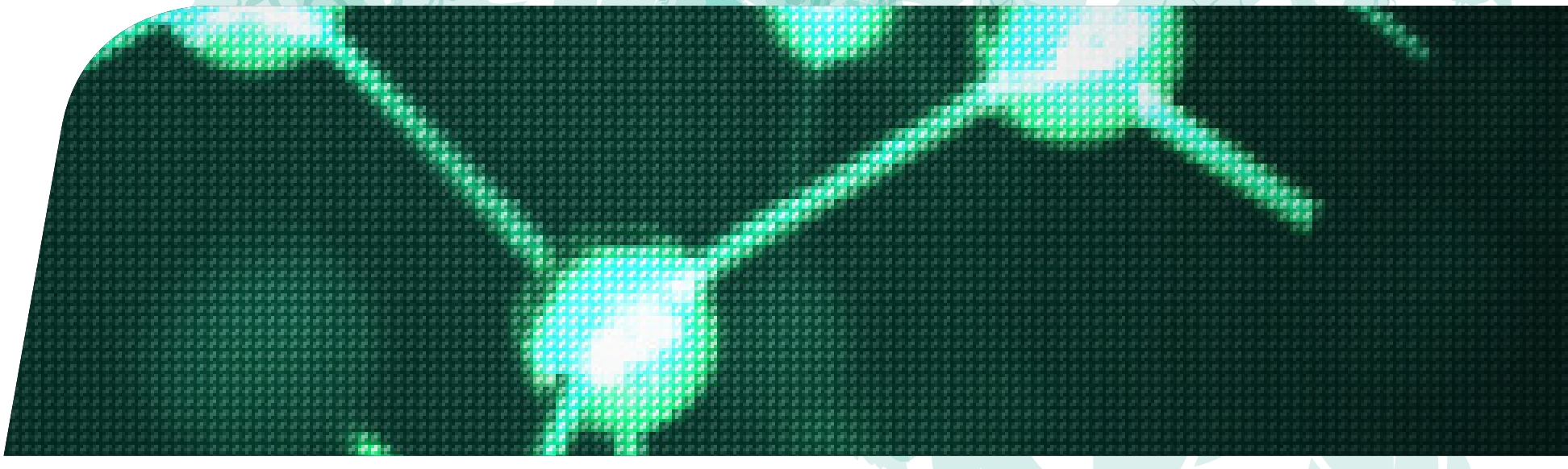
Science
Comparative analysis of bat genomes provides insight into the evolution of flight and immunity.

Science
Genomic diversity and evolution of the head crest in the rock pigeon.

Nature
Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation.


4
I. DNA-level Research
5
1. De novo sequencing
8
2. Resequencing
10
3. Epigenome sequencing
12
II. RNA-level Research
13
1. RNA sequencing for transcriptome
13
2. RNA sequencing for quantification
13
3. Digital Gene Expression (DGE): tag profiling
13
4. Small RNA sequencing
13
5. Degradome sequencing
16
III. Protein-level Research
17
1. Proteome profiling
17
2. Quantitative proteomics
17
3. Modification proteomics
17
4. Target proteomics

Contents



Multi-level Sequencing Technologies for Plant and Animal Research

Sequencing technologies have revolutionized plant and animal research, allowing us to explore each species from DNA, RNA and protein levels (Fig. 1). Equipped with cutting edge technologies, tremendous computing power, and strong bioinformatics know-how, we strive to help our customers and collaborators to accelerate their scientific research and molecular breeding efforts.

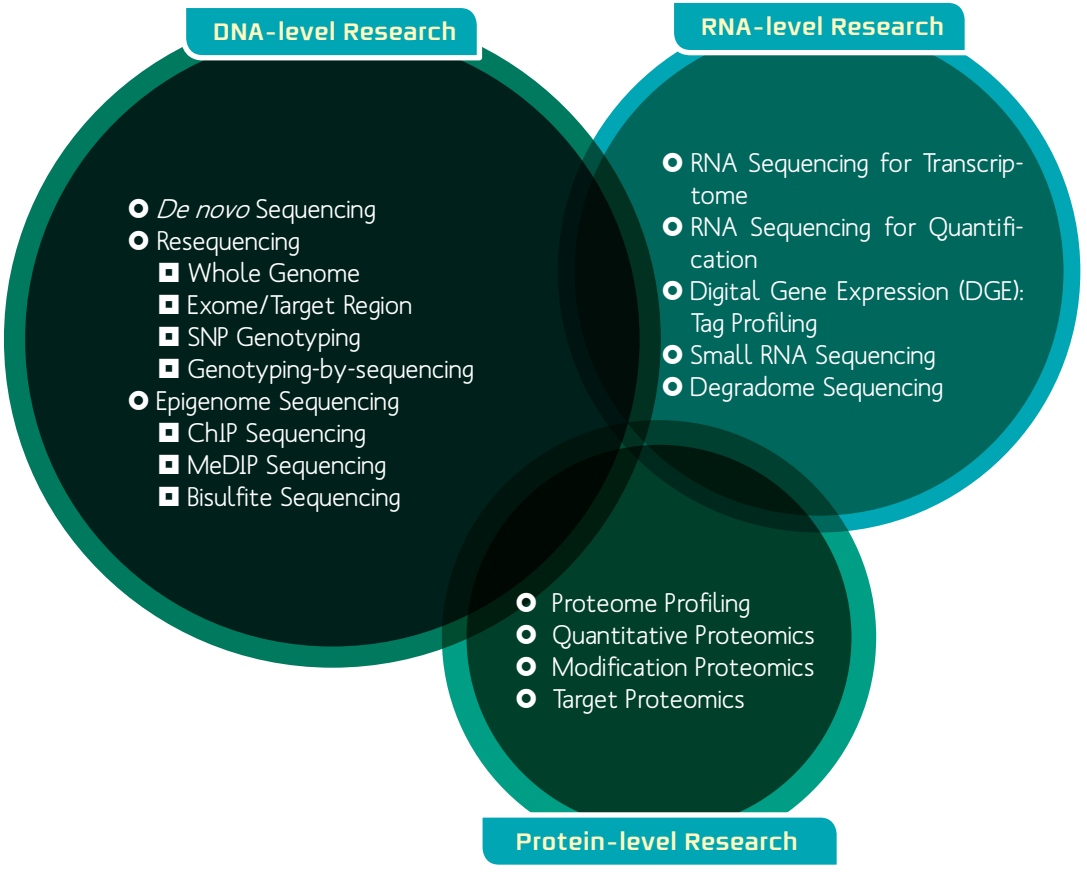


Fig.1. Multi-level sequencing technologies at BGI

I. DNA-level Research

DNA-level research includes efforts on genome-wide studies to determine the entire DNA sequence of an organism (*de novo* sequencing), fine-scale genetic variation analysis (resequencing), and epigenomics studies (Fig. 2). These approaches are applied more and more in a variety of biological research areas as well as in accelerating the process of molecular breeding.

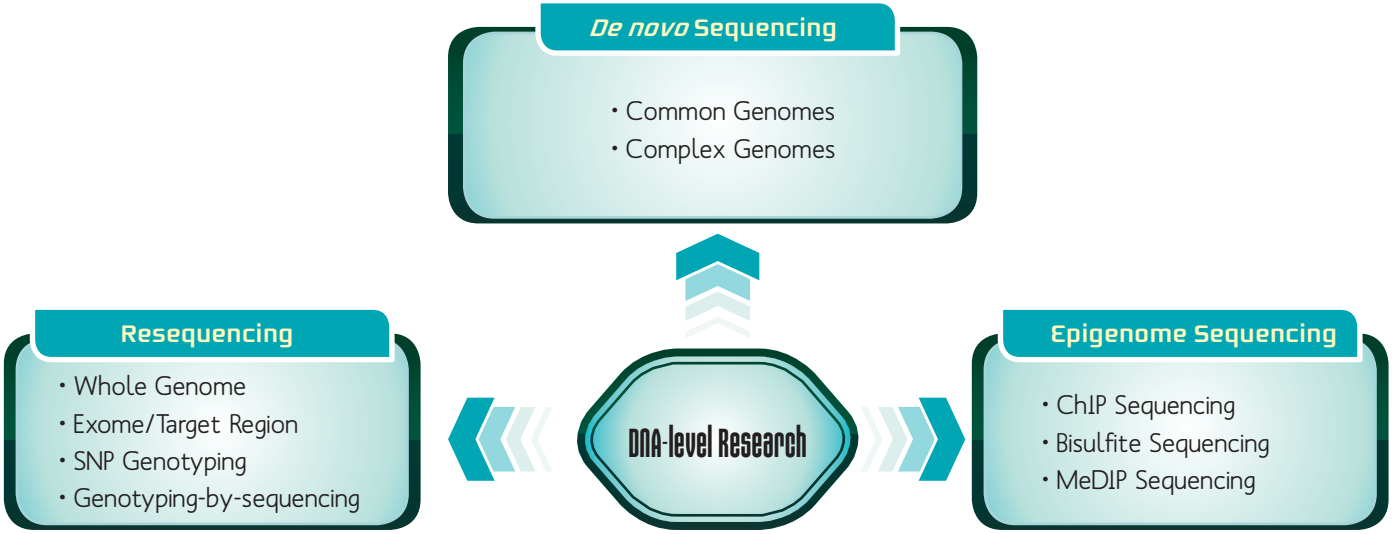


Fig. 2. DNA-level research at BGI

1. *De novo* sequencing

De novo sequencing aims to sequence a species afresh from the beginning without referencing any previous sequencing data of the species. Decoding the whole genome sequence is one of the most effective methods to understanding a species. As shown in Table 1, for common genome, we use whole genome shotgun (WGS) strategy to obtain the whole genome map. For complex genome, two different strategies are used to obtain the whole genome map: (1) WGS+BAC to BAC/fosmid to fosmid, and (2) ultra-deep de novo strategy (Fig 3).

Table 1. Genome characteristics for common and complex genomes

Genome characteristics	Common genome *	Complex genome **
Ploidy	Haploid or homozygous diploid	Heterozygous diploid or polyploid
Heterozygosity rate	< 0.5%	> 0.5%
GC content	35% ~ 65%	< 35% or > 65%
Repeats content	< 50%	> 50%

*All characteristics are included. ** At lease one characteristic is included.

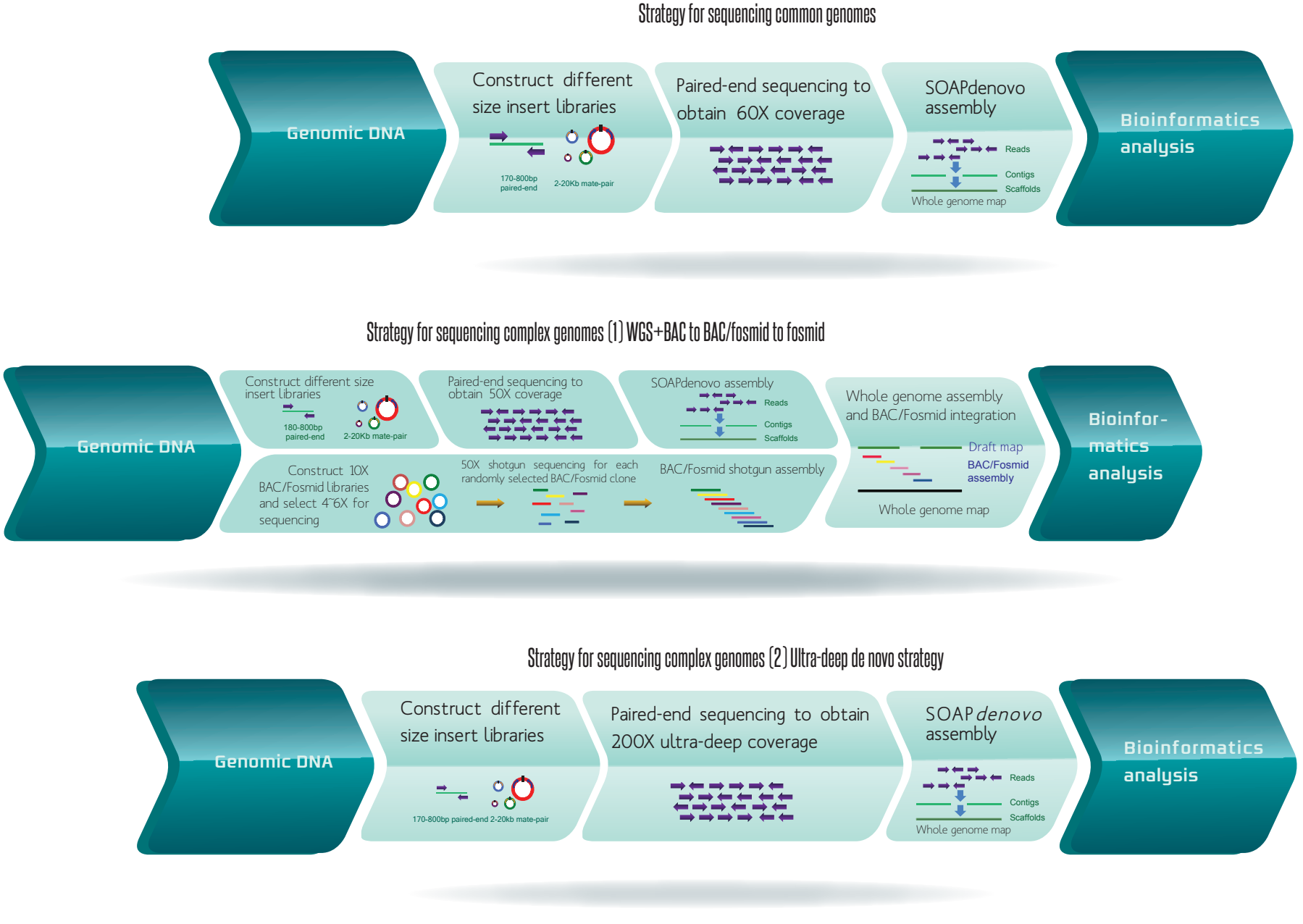


Figure 3: Technology pipelines for *de novo* sequencing

Bioinformatics analysis for de novo sequencing

Technology	Bioinformatics Analysis
De novo Sequencing	**Genome annotation:** 1. Repeat sequence (mainly tandem repeats and interspersed repeats) annotation 2. Gene structure prediction, including genes, CDS sequences and protein sequences 3. Gene function annotation using blastp to SwissProt and TrEMBL databases 4. ncRNA annotation, including tRNA, rRNA, snRNA and miRNA, etc.
	Evolution analysis: 1. Analysis of orthologous gene clusters using single copy gene families and multi-gene families 2. Construct a phylogenetic tree using single-copy orthologuous genes 3. Comparative analysis of whole genome alignment 4. Analysis of segmental duplication ranging in size from 1 Kb to several Mb 5. Detect conserved element of coding sequences, regulatory elements or other functional elements

2. Resequencing

With the reference genome sequences available for many organisms, resequencing the whole genome or target regions of an individual or population is fundamental for identifying genetic variations and understanding its biological significance.

1) Whole genome resequencing

DNA sequencing can be used to gain important information on genes, genetic variations and gene functions for biological studies. The growing collection of publicly available reference genome sequences will provide a foundation for whole genome resequencing studies.

2) Exome / Target region sequencing

New generation capture technology allows deep sequencing of exome or target regions, with more accuracy, cost-efficiency and effectiveness. Specifically, coding sequences and traits-related QTLs can now be studied. BGI provides two capture platforms:

- NimbleGen capture array
- Agilent SureSelect system

3) SNP genotyping

Sequencing is a powerful tool for known and novel SNP detection. For known SNPs, genotyping offers a cost-effective way to validate and screen a large number of samples. BGI provides genotyping services by employing Illumina BeadArray technology. Many types of plant and animal genotyping chips are available such as porcine, bovine, ovine, canine, equine, and maize. In addition, customized chips can be designed upon request.

4) Genotyping-by-sequencing

Genotyping by sequencing(GBS) is a simple and cost-effective genotyping and mapping method based on next-generation sequencing of genomics regions targeted by restriction egzymes (REs). By choosing the right REs, repetitive regions can be avoided and lower copy regions can be targeted with higher efficiency, which greatly simplifies the computationally challenging alignment for genetically diverse species. GBS can be used effectively in large scale SNP identification and for exploring intra-species diversity, constructing haplotype maps and performing genome-wide association studies (GWAS). GBS is especially useful for genotyping and mapping genetically diverse and large genome species, with or without a reference genome.

Bioinformatics analysis for resequencing

Technology	Bioinformatics Analysis
Whole Genome Resequencing	**Individual:** 1. Assembly of consensus sequences 2. SNP, InDel and Structure Variaton (SV) detection, annotation and statistics
	Population: 1. SNP, InDel and Structure Variaton (SV) detection, annotation and statistics 2. Linkage Disequilibrium (LD) analysis 3. Phylogeny tree analysis 4. Population structure analysis 5. Principal Component Analysis (PCA) 6. Polymorphisms analysis
Exome / Target Region Sequencing	**Individual:** 1. SNP, InDel detection, annotation and statistics
	Population: 1. SNP, InDel detection, annotation and statistics 2. Population SNP calling 3. Haploview: linkage disequilibrium and haplotype prediction 4. Positive selection signals detection
SNP Genotyping	1. SNP calling, sample level QC and SNP level QC 2. Imputation, population structure and association analysis, SNP annotation
Genotyping-by-sequencing	1.SNP detection, annotation, and statistics 2.Phylogeny tree analysis; Population structure analysis; Principal Component Analysis (PCA) 3.Genotyping, Genetic map construction, QTL mapping analysis (with phenotype data), Integration of genetic map 4.GWAS

3. Epigenome sequencing

Epigenomics is the study of heritable changes other than those encoded in DNA sequence. Cytosine methylation of DNA is the most well-studied epigenetic phenomenon, although epigenetic changes also encompass non-DNA methylation mechanisms, such as covalent histone modifications, micro-RNA interactions, and chromatin remodeling complexes. Epigenomics research is growing tremendously through next generation sequencing and innovative molecular and computational techniques.

1) ChIP sequencing

ChIP sequencing combines chromatin immunoprecipitation (ChIP) with next generation sequencing technology to identify the binding sites of DNA-associated proteins. It is widely applied to the study of histone modification and transcription factors regulation.

2) Bisulfite sequencing

Bisulfite sequencing is the gold standard for DNA methylation analysis. By combining bisulfite treatment and high-throughput sequencing, it generates high-accuracy methylome and provides a reference for differential DNA methylation analysis across a large number of samples.

3) MeDIP sequencing

MeDIP sequencing is a cost-effective method to study whole genome DNA methylation based on immunoprecipitation. It efficiently compares DNA methylation modification patterns between different samples.

Bioinformatics analysis for epigenome sequencing

Technology	Bioinformatics Analysis
ChIP Sequencing	1. Reads alignment 2. Genome-wide distribution of ChIP sequencing reads 3. Peak scanning and distribution 4. Peak-related genes scanning and gene ontology (GO) function analysis 5. Difference analysis of multi-samples 6. UCSC genome browser instruction 7. ChIP sequencing reads distribution around TSS 8. Motif analysis
Bisulfite Sequencing	1. Reads alignment 2. Sequence depth and coverage analysis 3. Calculation of methylation level 4. Global trends of methylome 5. Genome-wide methylation profiling 6. Identification of differentially methylated regions (DMRs)
MeDIP Sequencing	1. Reads alignment 2. Distribution of uniquely mapped reads 3. Peak scanning and genome-wide distribution 4. Difference analysis of multi-samples 5. Combined analysis of MeDIP sequencing with RNA sequencing data

II. RNA-level Research

Compared with DNA, RNA is more dynamic, as it reflects the functional status of specific cells under a particular condition. RNA-level research provides tools that help researchers gain a better understanding of how genes and pathways are involved in biological processes (Fig. 4).

Fig. 4. RNA-level research at BGI

1. RNA sequencing for transcriptome

RNA sequencing for transcriptome analysis provides efficient ways to precisely quantify transcript levels, confirm or revise previously annotated genes, map exon/intron boundaries, detect post-transcriptional mutations, and identify gene fusions.

2. RNA sequencing for quantification

RNA sequencing for quantification analysis is used to analyze gene expression of certain biological objects under specific conditions.

3. Digital Gene Expression (DGE) tag profiling

Digital Gene Expression (DGE) tag profiling offers unparalleled depth, specificity, and sensitivity for confident novel and rare transcripts discovery in different tissues under specific conditions.

4. Small RNA sequencing

Small RNA sequencing is the most reliable method to identify and profile small RNAs, providing valuable insights into the mechanisms involved in gene regulation.

5. Degradome sequencing

Degradome sequencing is a novel method for profiling miRNAs mediated cleavage events in organisms, providing valuable insights in microRNA research.

Bioinformatics analysis for RNA sequencing

Technology	Bioinformatics Analysis
RNA Sequencing for Transcriptome	*De novo:* 1. Unigene function annotation, gene ontology (GO) classification and metabolic pathway analysis 2. Protein coding region prediction (CDS) 3. Identification, gene ontology (GO) classification and pathway enrichment analysis of differentially expressed unigene
	Resequencing: 1. Gene expression, annotation and analysis of differentially expressed gene (DEGs) 2. Refinement of gene structures and identification of alternative spliced transcripts 3. Predication of novel transcripts 4. SNP analysis
RNA Sequencing for Quantification	1. Statistics of gene expression 2. Expression pattern analysis of differentially expressed genes (DEGs) 3. Gene ontology (GO) and pathway enrichment analysis of differentially expressed genes (DEGs) 4. Protein-protein interaction network analysis

Technology	Bioinformatics Analysis
Digital Gene Expression (DGE): Tag Profiling	1. Gene expression annotation 2. Screening of differentially expressed genes (DEGs) 3. Anti-sense transcripts annotation 4. Novel transcripts detection 5. Expression pattern analysis of differentially expressed genes (DEGs) 6. Gene ontology (GO) and pathway enrichment analysis of differentially expressed genes (DEGs) 7. Protein-protein interaction network analysis
Small RNA Sequencing	1. Common and specific sequences between samples 2. Analysis results of SOAP (map to genome) 3. Annotation results (Align small RNA reads to miRBase/repeat/Rfam/Genbank/Exons and introns) 4. Novel miRNA prediction 5. Analysis of differentially expressed miRNA 6. Cluster analysis 7. Family analysis of known miRNA 8. miRNA base edit analysis 9. miRNA target prediction 10. Function analysis of miRNA target genes
Degradome Sequencing	1. Common and specific sequences between samples 2. Analysis results of SOAP (map to genome) 3. Remove polyN reads 4. Annotation results (Align small RNA reads to Rfam/Genbank/Exons and introns) 5. Identify mRNA degradation 6. Count the expression of each cutting site 7. Identify mRNA-miRNA pairs

III. Protein-level Research

Proteomics research is the study of the structure, function and regulation of proteins, the principal and vital constituents of living organisms, as well as the main components of the physiological metabolic pathways of all cells. Proteomics can identify the presence of proteins, detect post-translational modifications, and provide quantification measurement (Fig. 5).

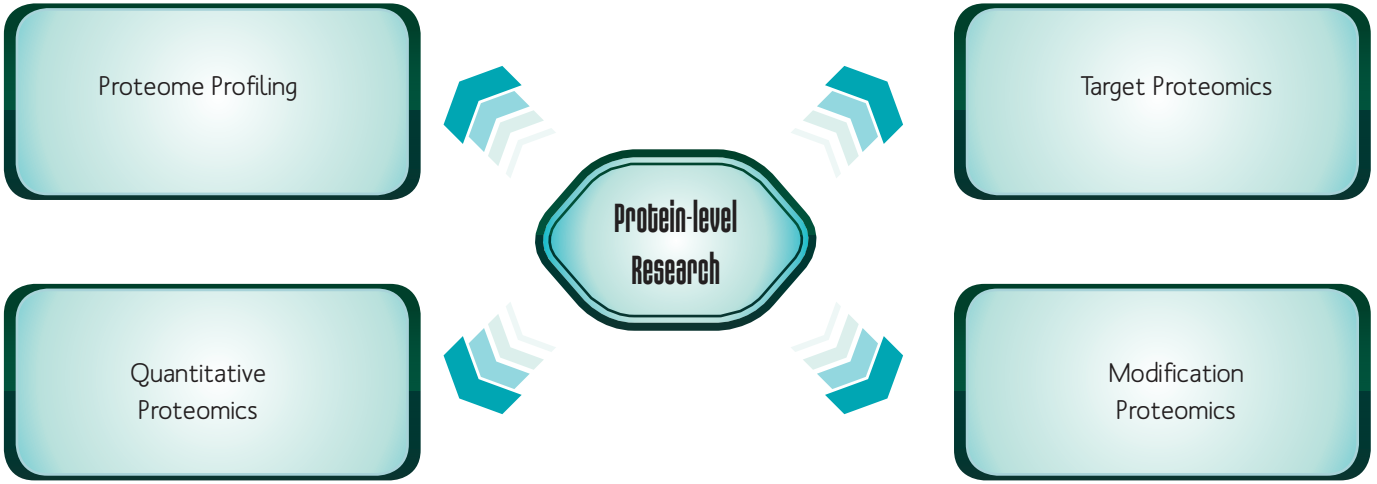
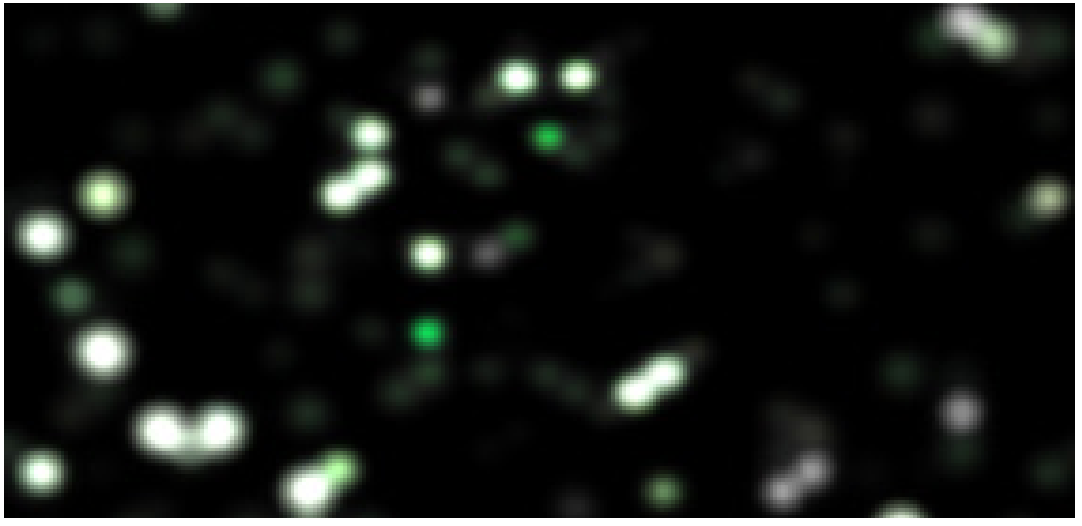


Fig. 5. Protein-level research at BGI



1. Proteome profiling

The study of protein profiling aims to launch the maximum amount of protein identification from the complex sample extract, including tissue, plasma, and body fluid It can provide a fundamental knowledge for proteomics analysis.

2. Quantitative proteomics

Quantitative proteomics aims to obtain quantitative information about all proteins in a sample, and functional information of differentially expressed proteins and differentially expressed proteins between samples.

3. Modification proteomics

Modification proteomics aims to identify the type and site of modified proteins, classify the functions, and analyze the signal transduction pathways of modified proteins.

4. Target proteomics

Target proteomics is used in the verification of global proteomics data, the discovery of lower abundance proteins, detection of protein post-translational modifications, discrimination of highly homologous protein isoforms, and validation of biomarkers.

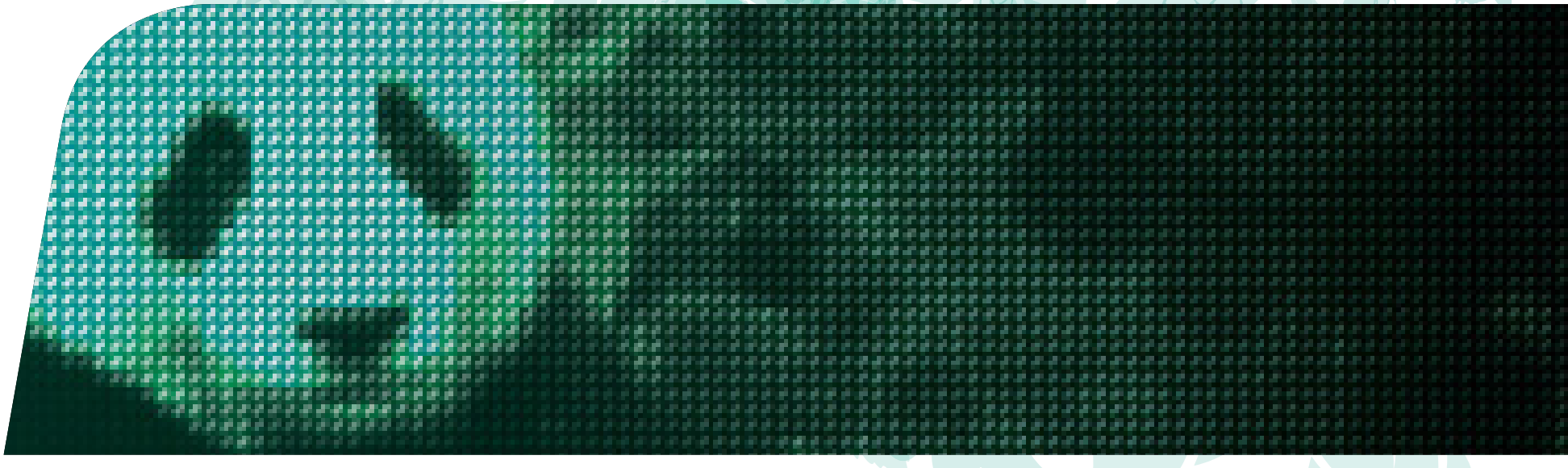
Bioinformatics analysis for protein-level research

Technology	Bioinformatics Analysis
Proteome Profiling	1. Protein identification 2. Protein functional annotation 3. Protein gene ontology (GO) classification 4. Protein pathway analysis
Quantitative Proteomics	1. Protein identification 2. Protein quantification 3. Protein functional annotation 4. Protein gene ontology (GO) classification 5. Gene ontology (GO) enrichment analysis of distinct proteins 6. Protein pathway analysis
Modification Proteomics	1. Protein identification 2. Protein modification site identification 3. Protein functional annotation 4. Protein gene ontology (GO) classification 5. Protein pathway analysis
Target Proteomics	1. Transition list design 2. Target protein identification and quantification 3. Protein functional annotation 4. Protein gene ontology (GO) classification 5. Protein pathway analysis

Proteomics Mass Spectrometry Platform:

Customized analysis is available to meet your specific requirements. Please contact us to know more.

Building on high-throughput sequencing and a powerful computing platform, we will support your research to deliver biologically relevant results with unrivaled speed, accuracy and flexibility.



Applications
in Plant and Animal Research

I. Sequencing to reveal evolutionary process

De novo sequencing enables you to study genomes within the same species (Species Pan Genomes) or decode genomes in the same family or genus (Clade Genomes). See Figure 6 for evolution research based on *de novo* sequencing.

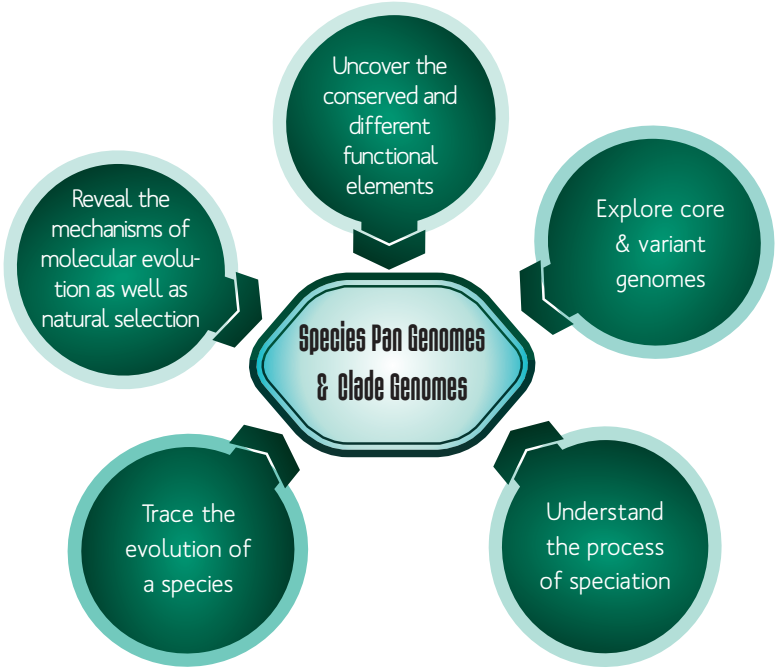
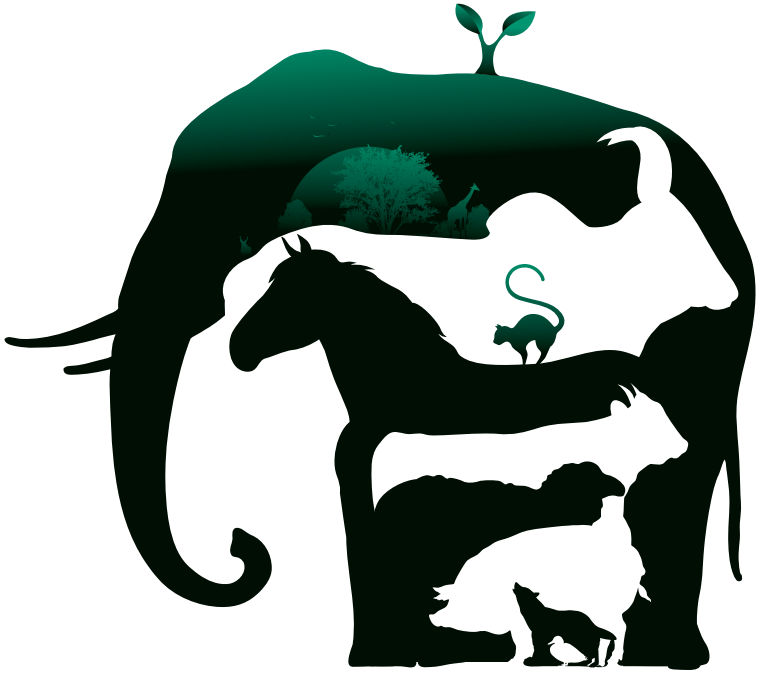


Fig. 6. Evolution research based on *de novo* sequencing



Bioinformatics analysis for species pan genomes and clade genomes:

1. Predict genomic functional elements and obtain differentiation characteristics
2. Detect gene family expansion or contraction and the birth or death of genes
3. Construct gene relationships
4. Analyze natural selection and biological adaptation
5. Reveal species divergence mechanism and progress
6. Detect genomic variation between species (InDels, duplication, retrotransposons)

II. Sequencing to accelerate molecular breeding

Molecular breeding includes a variety of applications, such as functional gene or marker detection, functional characterization of genes and new varieties development. BGI provides various solutions for selected breeding applications including gene/marker discovery and development of new varieties.(Fig. 7).

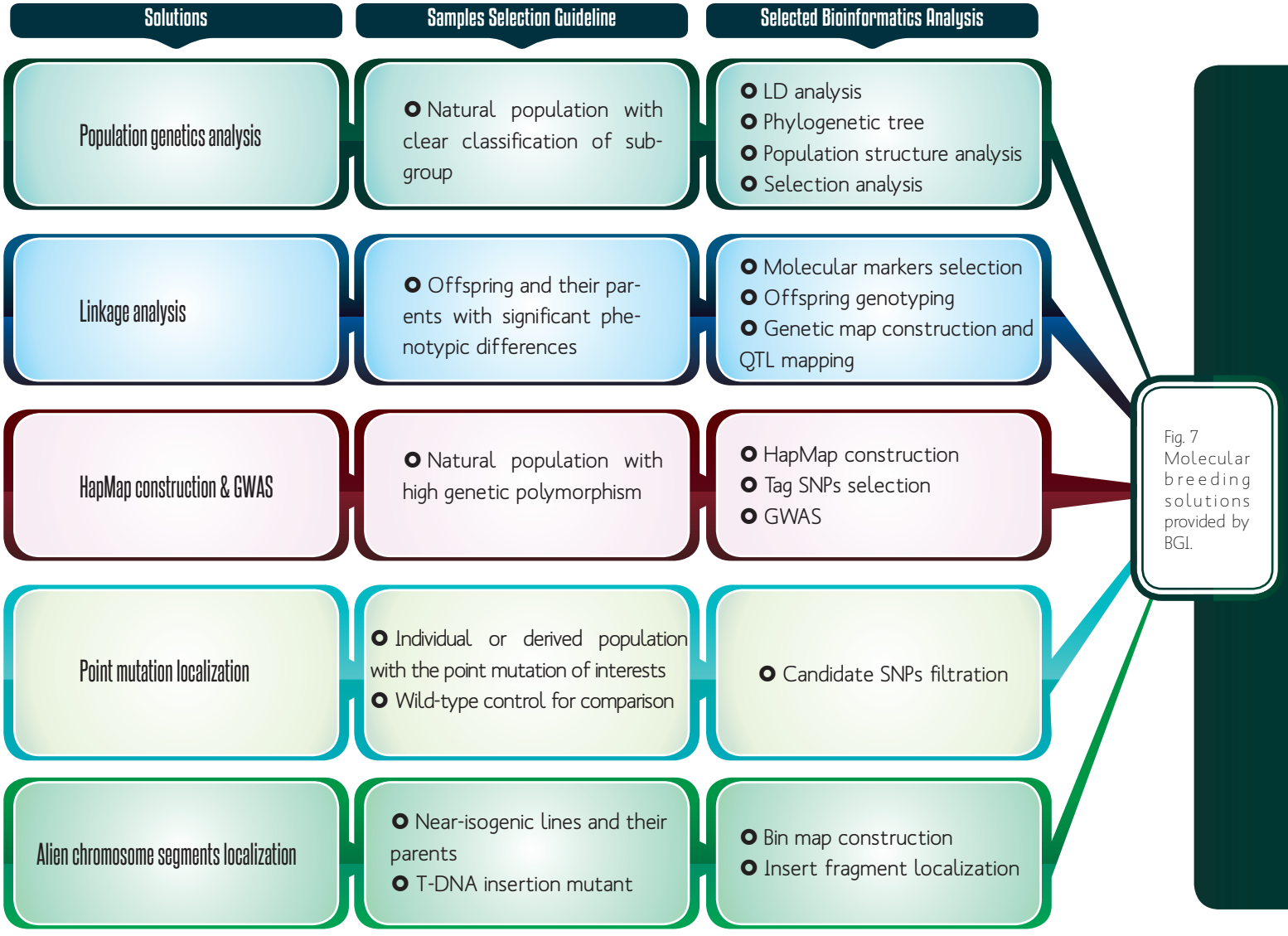


Fig. 7 Molecular breeding solutions provided by BGI.

Population Genetics Analysis

Representative varieties such as domesticated and wild lines could be resequenced to reveal the mechanism of domestication event and to detect candidate genes which may be strongly selected during the domestication.
In addition, for germplasm, population genetics analysis identifies the evolutionary relationships between different materials, which is very helpful for heterosis breeding.

Linkage Analysis

Linkage analysis is one of the most important strategies in molecular breeding. BGI provides the following services: molecular markers discovery, offspring genotyping, genetic map construction, and QTL fine mapping. Common group types for linkage analysis include F2, BC, RIL, and DH populations of targeted species.

HapMap Construction and GWAS

Haplotype, a set of closely linked genetic markers on a chromosome that tend to be inherited together, can be identified by patterns of SNPs through resequencing. A haplotype map (HapMap) can be used to identify genetic variations of inherited traits by Genome Wide Association Studies (GWAS).

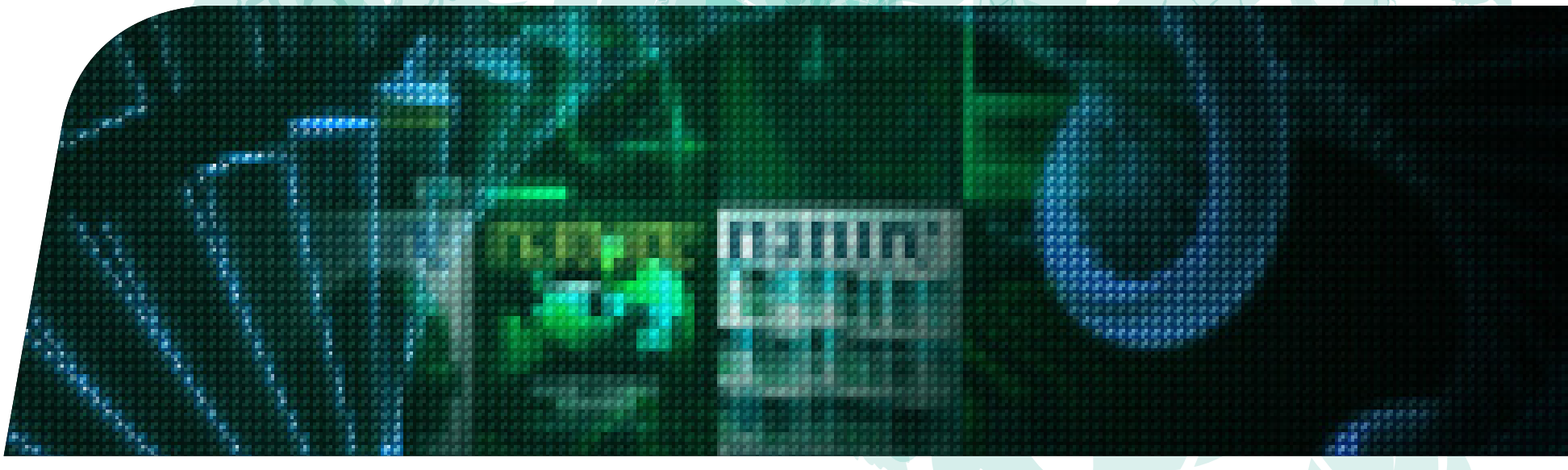
Point Mutation Localization

Fine mapping of point mutations is important to functional gene discovery. It can be quickly achieved by resequencing the mutants, which are usually induced by chemical mutagens such as EMS or ENU, or its derived population such as F2 and BC1 progeny.

Alien Chromosome Segments Localization

The substitution or insertion of alien chromosome segments often occurred during backcross or T-DNA insertion. With next generation sequencing, we could rapidly obtain the sequence of substitution segments or the location of insertion segments. With this information, functional genes that caused phenotype variation can be easily detected.


BGI



Selected Publications from BGI in Plant and Animal Research

1. The Cucumber Genome

The genome of the cucumber, *Cucumis sativus* L. *Nature Genetics.* 2009. 41: 1275-1281.

The cucumber genome provides insights into the genetics of the whole cucurbit family, which includes pumpkins, squash, melon and watermelon. It also offers insights into traits such as vascular transport systems, disease and pest resistance, the "fresh green" odor of the fruit, bitter flavors, sex expression and how the plants form flowers and fruits.

2.The Potato Genome

Genome sequence and analysis of the tuber crop potato *Nature*. 2011.475:189-195.

This potato genome(844Mb) marks an important milestone in potato genome research, revealing new insights into the evolutionary history of the potato genome, causes of inbreeding depression, and potential mechanisms of tuber initiation and development. These insights will genetate great interest among botanists and breeders worldwide and will facilitate the genetic engineering of this vital crop.

3.The Pigeonpea Genome

Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers.*Nature Biotechnology* 2011. 30: 83-89.

The pigeonpea genome (833M) facilitates the identification of the genetic basis of agronomically important traits, and accelerates the development of improved pigeonpea varieties that could improve food security in many developing countries.

4. The Cotton Genome

The draft genome of a diploid cotton Gossypium raimondii. *Nature Genetics*.2012. 44: 1098–1103.

The cotton genome (775M) provides an invaluable resource for the study and genetic improvement of cotton quality and output, and sheds new light on understanding the genetic characteristics and evolutionary mechanisms underlying cotton and its close relatives.

5.The Wheat A-genome

Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature*. 2013. 496: 87–90.

The Triticum urartu (AA) draft genome sequence (4.94 Gb) provides a diploid reference for analysis of polyploid wheat genomes and is a valuable resource for the evolution, domestication, and genetic improvement of wheat.

6. The Panda Genome

The sequence and *de novo* assembly of the giant panda genome. *Nature*. 2010. 463: 311-317.

The Panda genome is the first genome completed entirely through the use of a next generation sequencing platform. This gives clues to understand everything from a Panda's strict bamboo diet to its genetic diversity. In the future it may also aid in the conservation of Pandas.

7. The Ant Genome

Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science*. 2010. 329: 1068-1071.
The ant genomes provide insight into epigenetics and longevity. This makes ants a new genetic model organism to study the impact of epigenetics on everything from aging, developmental, reproductive and behavioral plasticity, to metabolism and social behaviors.

7.The Oyster Genome

The oyster genome reveals stress adaptation and complexity of shell formation.Nature.2012.490:49-54.
The pig genome (2600Mb) sequence illuminates the evolution of Sus scrofa and confirms its speciation in South East Asia and subsequent domestication in multiple regions across Eurasia. The high-quality annotated reference genome sequence has already proven to be a critical framework for comparing individual genomes.

9.The Pig Genome

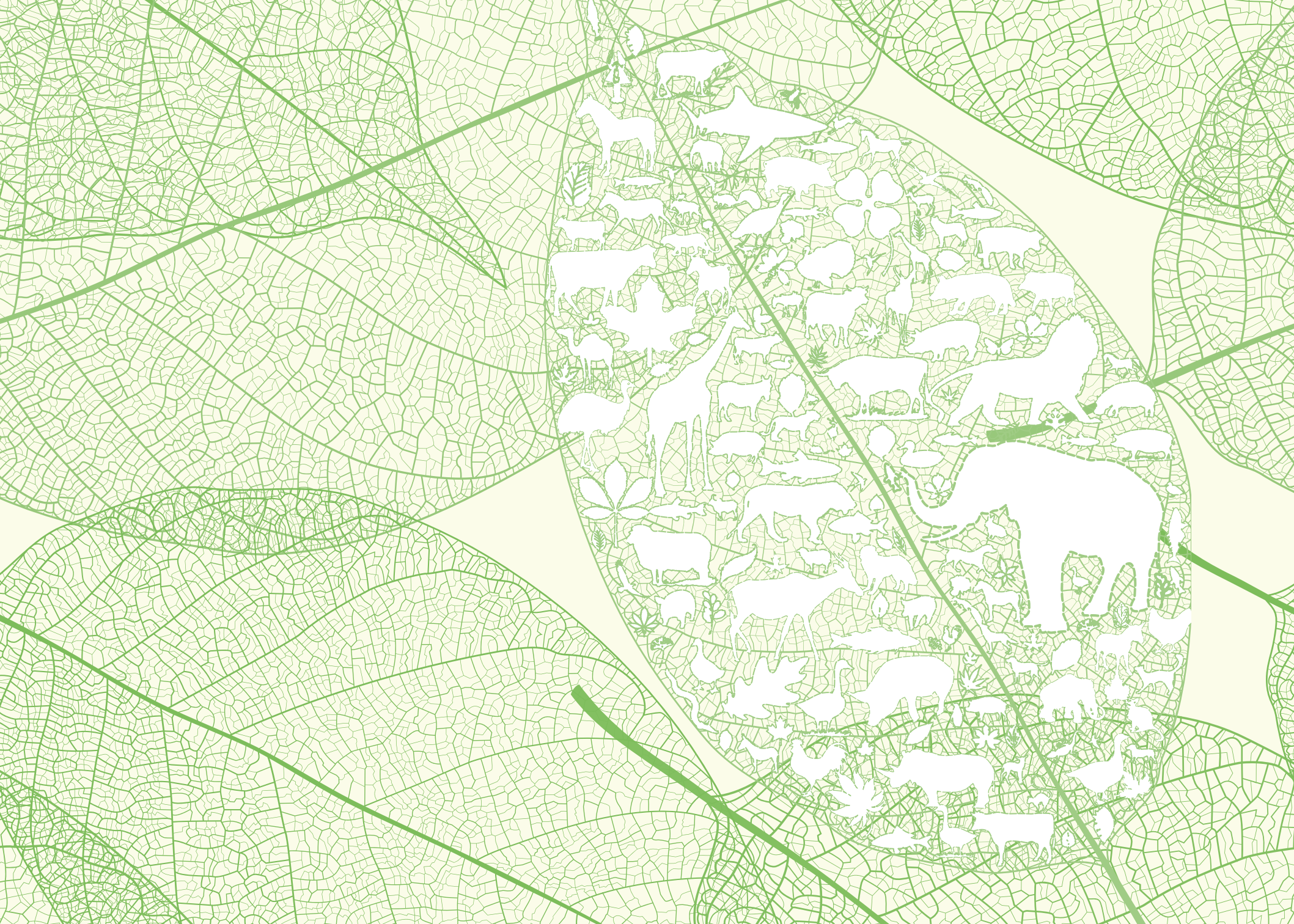
Analyses of pig genomes provide insight into porcine demography and evolution.*Nature*.2012. 491:393-398.

This research generated the first genome-wide DNA methylation map for a bird species. Generally, chickens display analogous methylation patterns with that of animals and plants. These DNA methylome maps will be useful for futher studies on epigenetic gene regulation in chickens and other birds.

10.The Pigeon Genome

Genomic diversity and evolution of the head crest in the rock pigeon. Science. 2013. 339: 1063-1067.

The domestic rock pigeon (*Columba livia*) genome (1.3 Gb) is a promising model with which to explore the genetic architecture of derived, constructive phenotypes in a bird that is amenable to genetic, genomic, and developmental investigation.



Contact Us

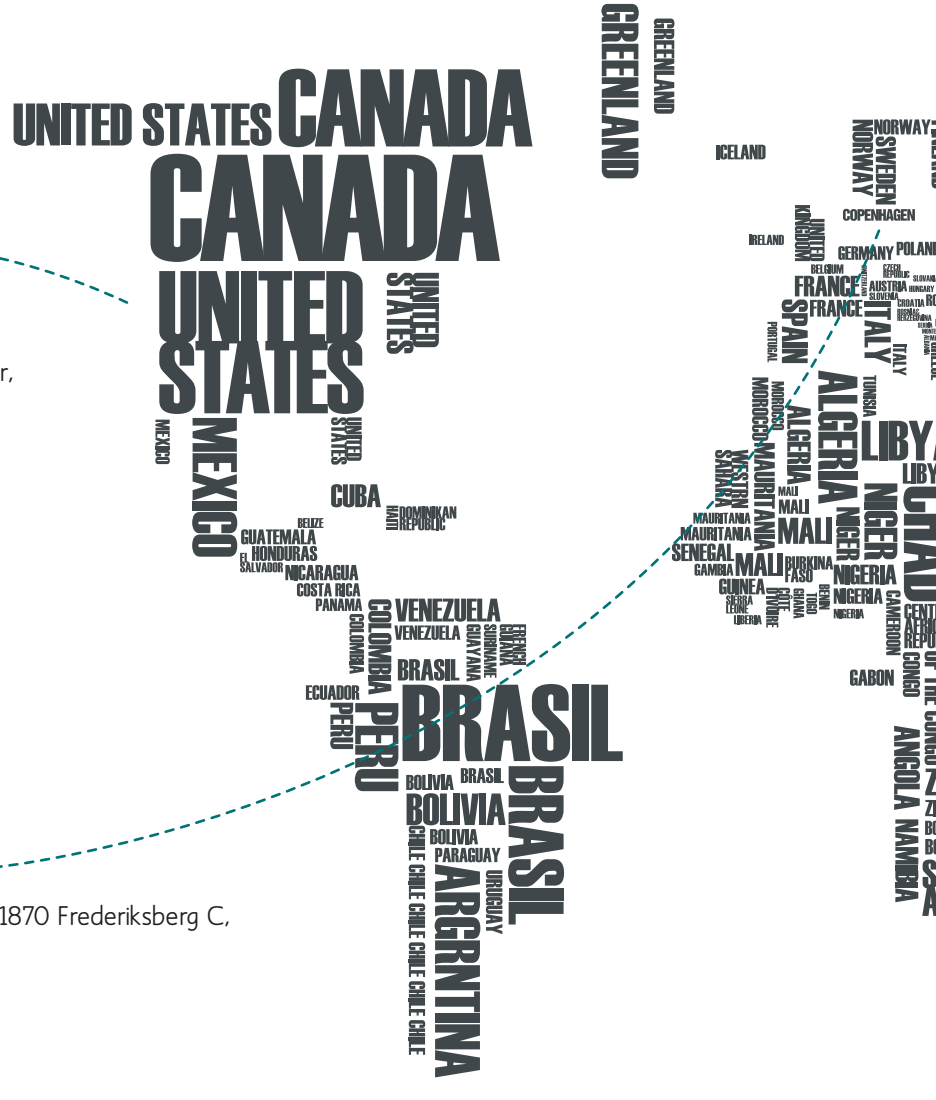
Offices & Locations

North America (Boston)

BGI Americas Corporation, One Broadway, 14th Floor, Cambridge, MA 02142, USA
Tel: +1-617-500-2741
Email: info@bgiamericas.com
www.bgiamericas.com
www.bgitechsolutions.com

Europe (Copenhagen)

BGI Europe A/S, Bülowsvej 15, DK-1870 Frederiksberg C, Copenhagen, Denmark
Tel: +45-7026-0806
Email: bgieurope@genomics.cn
www.bgitechsolutions.com



China (Mainland)

BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen, 518083, China
Tel: 400-706-6615
+86-755-25273620
Email: info@bgitechsolutions.com
www.bgitechsolutions.cn

Asia Pacific (Hong Kong)

BGI Hong Kong Co. Limited, 16th Dai Fu Street, Tai Po Industrial Estate, Tai Po, Hong Kong
Tel: +852-3610 3510
Email: bgihk.enquiry@genomics.org.cn
www.bgitechsolutions.com


华大科技
BGI tech
Premier Scientific Partner