BGI-Tech P&A Pan-genome Research Solution

Research Protocol

Contents

Product Background

1. Introduction

As research develops, the single species reference genome has been exposed to its limitations in the detection of genetic variation. This has directly led to the concept of the pan-genome.

A species can be described by its “pan-genome” (pan, from the Greek word π αν, meaning whole), which includes a core genome containing genes present in all genomes and a dispensable (variable/accessory) genome composed of genes absent from one or more genes that are unique to each genome.[1]

A single reference genome is not always sufficient to understand the genetic basis of different traits. For example, plants have many important agronomic trait genes that are often found in the dispensable genome. A single reference genome may not detect most of the presence/absence variants (PAV) and structure variants (SV). In this case, the value of constructing a pan-genome is immeasurable.

Today, the long-read sequencing technologies represented by Pacific Biosciences (PacBio) HiFi and Oxford Nanopore Technologies (ONT) have overcome the defects of low throughput of Sanger sequencing and read length of NGS sequencing technology. The continuous optimization combined with the assembly algorithm provides great convenience for the *De novo* assembly of genomes. The construction of pan-genome using long-read sequencing technology has been widely used in the study of animal and plant genomics to evaluate intra-species genetic diversity in a more comprehensive way. It also aids in the exploration of cross-species gene exchange, domestication, and improvement processes.

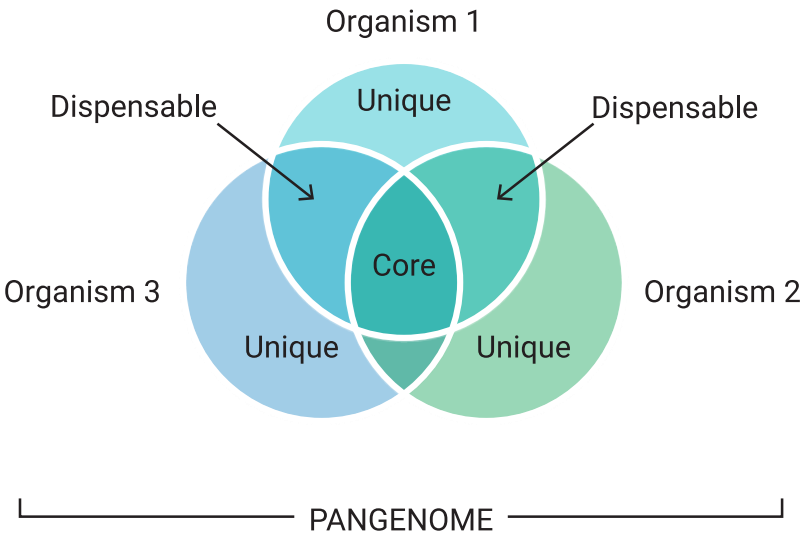


Figure 1. Example of a pan-genome. [2]

2. Case Study

Case 1. Extensive variation within the pan-genome of cultivated and wild sorghum.[3]

Sample collection: A widely representative sorghum pan-genome was constructed by *De novo* splicing of 13 varieties containing wild sorghum and cultivated sorghum, together with three previously published reference genomes of cultivated sorghum.

Sequencing strategy: PacBio long-read sequencing > 80X; Short-read long sequence > 100X; RNA-Seq 15.7Gb per sample.

Cultivated sorghum and its inter-fertile wild relatives constitute the primary gene pool for sorghum. Understanding and characterizing the diversity within this valuable resource is fundamental for its effective utilization in crop improvement. Here, we report the analysis of a sorghum pan-genome to explore genetic diversity within the sorghum primary gene pool. The researchers assembled 13 genomes representing cultivated sorghum and its wild relatives, and integrated them with 3 other published genomes to generate a pan-genome of 44,079 gene families with 222.6 Mb of new sequence identified. The pan-genome displays substantial gene-content variation, with 64% of gene families showing presence/absence variation among genomes.

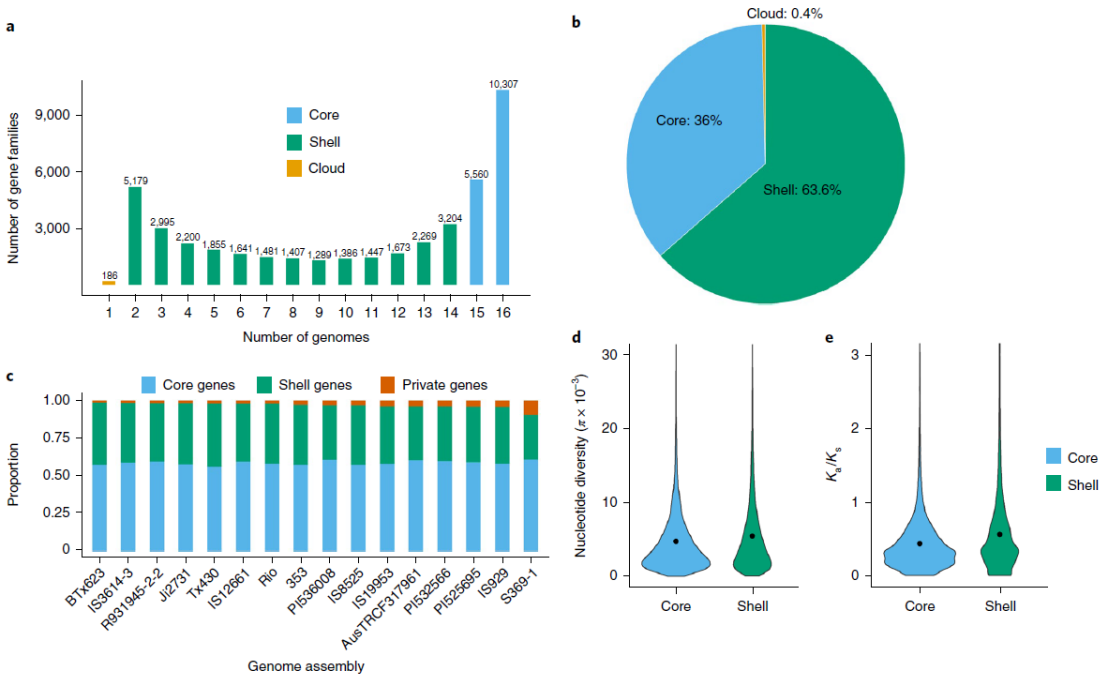


Figure 2. Sorghum pan-genome.

To investigate the effect of large segment structural variations on agronomic traits, genome-wide association analysis (GWAS) of grain color phenotypes was performed using genome-wide variation data from 839 cultivated sorghum varieties. On the significantly associated Yellow seed1 gene, which controls grain color, a PAV of 3216bp was identified in combination with pan-genomic data. Another candidate gene identified by GWAS, SbRC, is a homologue of the rice grain color gene Rc, which also has 416 bp PAV in the pan-genome. All of these PAVs had functional effects on gene structure and thus altered agronomic traits. The construction of sorghum pan-genome provides the basis for this method combining important agronomic traits GWAS with large fragment structural variation, and is expected to accelerate the research of sorghum functional genomics and breeding applications.

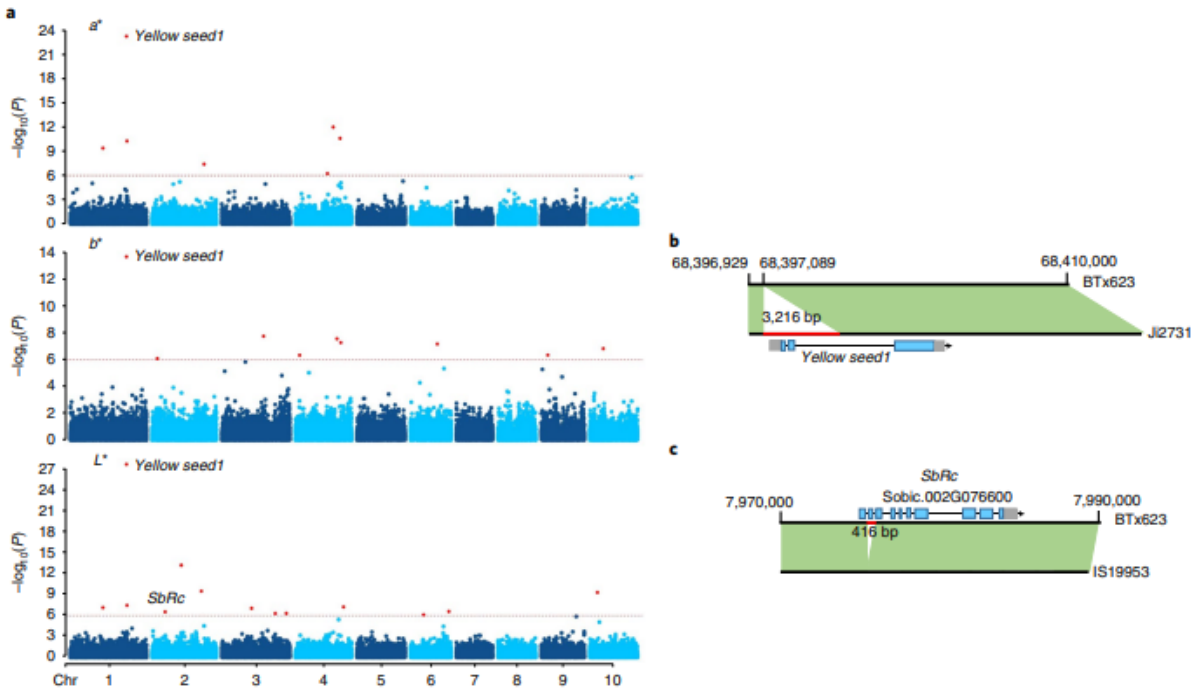


Figure 3. Presence/absence variation underlying grain-color variation in sorghum.

Case 2. High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation [4]

Sample collection: Researchers selected 1078 silkworm strains comprising 205 local strains, 194 improved varieties, and 632 genetic stocks of domestic silkworm (B. mori), and 47 wild silkworms (B. mandarina).

Sequencing Strategy: The average depth of short-read sequencing (DNBSEQ™ platform) is 65X; the depth of long-read sequencing (ONT platform) is 48X-277X.

The traditional breeding of silkworms has a long history and remarkable achievements, but it has hit a bottleneck since the 1990s. A systematic analysis of the genetic basis of domestication and selection for improvement is very important for breaking through the bottleneck of silkworm breeding.

The research team assembled 545 high-quality silkworm genomes, annotated 100 genomes, identified more than 43 million SNPs, 9.3 million Indel, 3.4 million SVs, and 7,308 new genes, and created a high-precision silkworm pan-genome map. The super pan-genome includes the most comprehensive genome information of the domesticated silkworm and the wild silkworm. It is the largest long-read long pan-genome with the largest sample size in the field of animals and plants in the world.

Figure 4. Phenotypic diversity in silkworms.

At the same time, the research team identified 468 domestication-associated genes and 189 improvement-associated genes, among which 264 and 185 were newly identified, respectively. The research team carried out in-depth research on various genetic variations, population structure, artificial selection, ecological adaptability, and economic traits of the silkworm, achieving rich innovating results. These genes will be important candidate targets for the molecular improvement of bombyx mori.

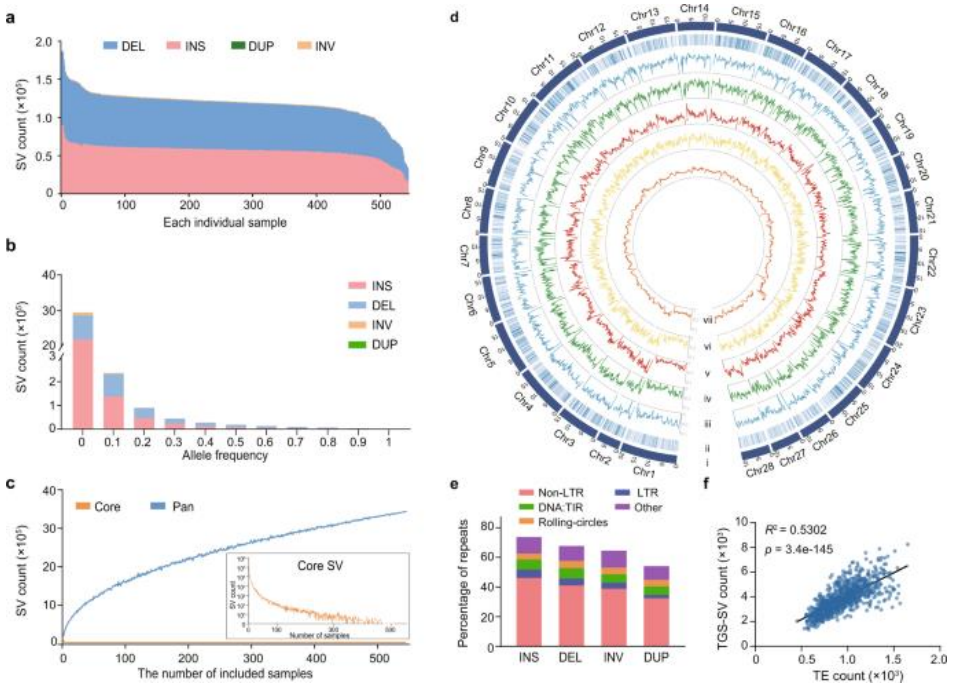


Figure 5. Characterization of SVs in 545 silkworm genomes.

Year Of Publication	Journal	Title	IF	Species	Research Strategy
2022	*Nature Communications*	High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation **(BGI)**	17.67	Silkworm	Deep resequencing on 1,078 silkworm species; ONT long-read sequencing on 545 representative silkworm species
2022	*Nature*	Genome evolution and diversity of wild and cultivated potatoes	49.962	Potatoes	44 potatoes PacBio (30X) + 7 of which had Hi-C data assembled to the chromosome level
2022	*Nature*	Graph pangenome captures missing heritability and empowers tomato breeding	49.962	Tomato	1 chromosome-level assembly (HiFi +Hi-C)+ 31 assembly (HiFi)+13 published genomes
2021	*Nature Plants*	Extensive variation within the pan-genome of cultivated and wild sorghum **(BGI)**	17.352	Sorghum	13 PacBio Sequel genomes +3 published genomes +839 published Darseq genotyping data
2021	*Nature*	A chickpea genetic variation map based on the sequencing of 3,366 genomes **(BGI)**	49.962	Chickpea	4 published genomes +3366 resequencing data
2021	*Cell*	Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations	41.582	Rice	31 PacBio genomes +2 published genomes +3010 published resequencing data
2020	*Cell*	Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato	41.582	Tomato	100 ONT pan-SV, and the genomes of 14 of them were assembled
2020	*Cell*	Pan-Genome of Wild and Cultivated Soybeans	41.582	Soybean	26 PacBio genomes +2898 resequencing data

Figure 6. Select publications of pan-genomic and research strategies.

In addition, with the declining sequencing cost, continuous progress of algorithms in recent years, and data accessibility of pan-genome research, the number of pan-genome articles published has increased. More and more species have been published, such as soybean, rape, barley, wheat, rice, sorghum, cotton, etc. In the field of animal and plant genomics, pan-genome can evaluate species' genetic diversity more comprehensively. The construction of species pan-genome has become a hotspot and the new frontier of genomics research.

Research Methods

Pan-genome assembly combines the advantages of a variety of sequencing technologies and provides reference sequence level genomes for further study of the origin, evolution, traits, and characteristics of species. At present, the published Pan-genome is mostly based on PacBio HiFi + ONT Ultra-long + Hi-C + NGS multiple sequencing strategies, which are characterized by the following:

PacBio HiFi: Because of the long read length and high quality of HiFi data, there is no error correction in the genome assembly, which allows for rapid genome assembly. It can be applied to different genome types where higher quality and continuity of genomes can be obtained in the assembly of high-repeat genomes, large complex genomes, allopolyploid genomes, and even autopolyploid genomes. For heterozygous genomes, two sets of haplotype genomes can be assembled. In terms of genome continuity, the Contig N50 of HIFI may be slightly lower than that of CLR mode because the read length is shorter than that of CLR, but its completeness is better than that of CLR.

ONT Ultra-long: This sequencing method can generate very long sequencing fragments that easily span the large repetitive regions in the genome. It can significantly improve the assembly effect of species genome and fill the gaps in the genome, making T2T (Telomere to Telomere) gap-free genome assembly possible.

Hi-C: Combining high-throughput sequencing technology with bioinformatics analysis methods, this approach studies the spatial position of the whole chromatin DNA on a genome-wide scale and obtain high-resolution three-dimensional chromatin structure information.

Pan-genome construction is mainly divided into the following ways [5]:

a) Alignment of reads from multiple samples to a reference is followed by assembly of unaligned reads into novel contigs. By adding these novel contigs to the original reference sequence, a pangenome reference can be constructed. Dispensable regions are determined based on mapping all reads back to the pangenome.

b) *De novo* assembly of the genomes of multiple accessions allows whole genome alignment approaches to identify dispensable genomic regions.

c) A pan-genome graph can be constructed from whole genome alignments or by *de novo* graph assembly, and efficiently stores variant information of dispensable regions as unique paths through the graph.

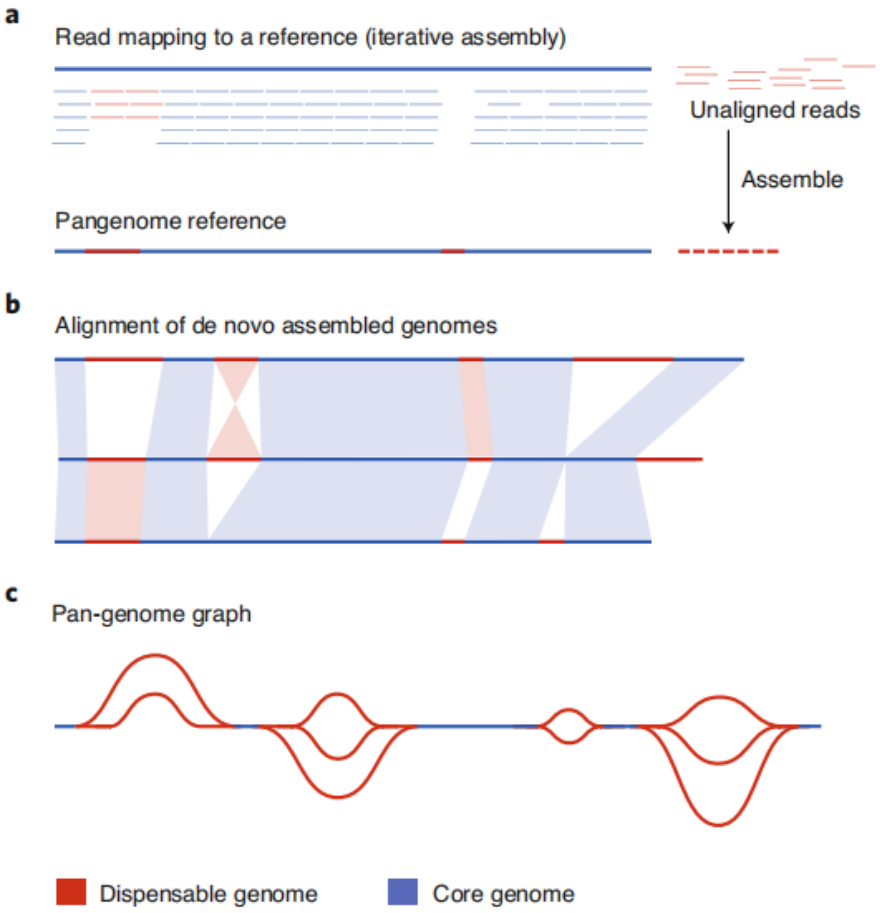
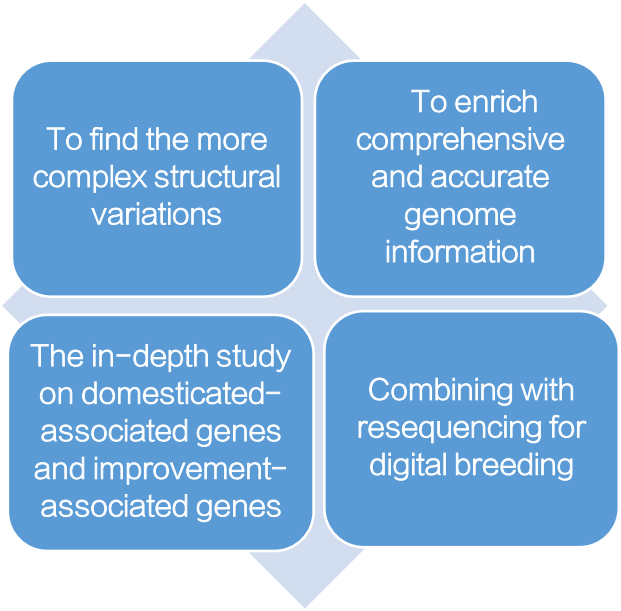


Figure 7. Comparison of pan-genome approaches.

Pan-genome research application features:



Reference Strategy

Samples Collection:

Sample selection is of great importance for Pan-genome research. BGI suggests that the number of representative samples should be ⩾10.

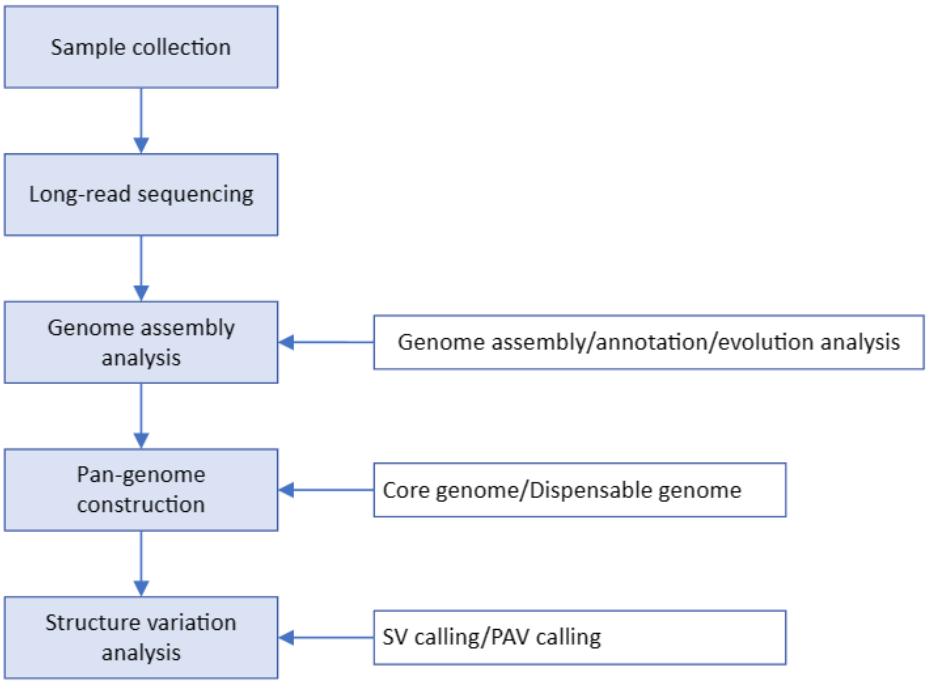


Figure 8. Strategy of Pan- genome research.

The pan-genomic analysis process of long read long sequencing is performed by:
1) The representative species were selected to complete the preliminary assembly of the genome using PacBio HiFi sequencing, and then the ONT Ultra-long data was used for "gap filling". Finally, the relative position information of genes on chromosomes was obtained by combining Hi-C technology to construct T2T genome. For other individuals, HiFi data was used to complete the preliminary assembly of the genome, and Hi-C data was combined to complete the assembly of the genome at the chromosome level.
2) Compare the genomes of different individuals, identify core genes and specific genes, and find out structural differences (SV, CNV, PAV, etc.). Combined with species resequencing data, GWAS analysis was performed on the traits of interest and key genes were located.

Therefore, for different species and different levels of complexity of the genome, we have introduced a variety of comprehensive solutions.

Sequencing Strategy:

Genome assembly usually includes DNBSEQ™-based genome survey, PacBio HiFi genome assembly, and Hi-C auxiliary assembly.

For the pan-genome construction, samples will be used for genome assembly, and it is recommended to select one sample for T2T assembly.

		General assembly	T2T assembly
Research Methods	**Sequencing Technology**	**Recommended Sequencing Depth**	**Recommended Sequencing Depth**
Genome Assembly	HiFi — PacBio	40-60X	40-60X
	Ultra-long — ONT	—	40 -100X
	Hi-C — DNBSEQ™	100X	100X

(The above steps are for reference only, Pan-genomic analysis and construction currently require customized evaluation)

Bioinformatics Analysis Content

Bioinformatics service content	
Genome Assembly (PacBio HiFi data+ Hi-C data)	1. Assembly 2. Assessment by short reads alignment 3. BUSCO assessment 4. Hi-C data auxiliary assembly
Gene Annotation	1. Repeat annotation 2. Gene structure annotation 3. Gene function Annotation 4. Transcription factors (plant)
Evolution	1. Gene family identification (Animal TreeFam; Plant OrthoMCL; ≤10 species) 2. Phylogenetic tree construction 3. Estimation of divergence time 4. Genome synteny analysis 5. Whole genome duplication analysis 6. Gene family expansion and contraction analysis
Pan-genome Analysis	1. Pan-genome construction 2. Core gene/ Dispensable gene analysis 3. SV/PAV detection and annotation 4. Graphical structure genome construction

References

[1] Tettelin, Hervé et al. “Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome".” Proceedings of the National Academy of Sciences of the United States of America vol. 102,39 (2005)

[2]https://www.pacb.com/blog/sequencing-101-looking-beyond-the-single-reference-genome-to-a-pangenome-for-every-species/

[3] Tao, Yongfu et al. “Extensive variation within the pan-genome of cultivated and wild sorghum.” Nature plants vol. 7,6 (2021)

[4] Tong, Xiaoling et al. “High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation.” Nature communications vol. 13,1 5619 (2022)

[5] Bayer, Philipp E et al. “Plant pan-genomes are the new reference.” Nature plants vol. 6,8 (2020)

www.bgi.com
info@bgi.com

BGI Americas

One Broadway, 14th Floor
Cambridge, MA 02142
U.S.A.

BGI Japan

Kobe KIMEC Center BLDG. 8F 1-5-2
Minatojima-minamimachi,
Chuo-ku, Kobe 650-0047 Japan

BGI Europe

Jutrzenki 12 A,
02-230 Warszawa,
Poland

BGI Australia

L6,CBCRC, 300 Heston Road,
erston, Brisbane,
Queensland 4006, Australia

BGI Asia

Building NO.7, BGI Park,
Yantian District Shenzhen,
Guangdong Province, China