Applications of Pangenome Research in Genetic Diversity and Functional Genomics (Postprint)

Authors: Kunli Xiang, He Wenchuang, Zou Yi, Peng Dan, Zhang Xiaoni, Liao Xuezhu, Wang Jie, Yang Jiankang, Wu Zhiqiang

Date: 2021-04-13T00:00:00+00:00

Abstract

Compared to single reference genomes that focus solely on mining individual genetic information, pangenome research can reflect the complete genetic information of an entire species or taxonomic group. With the continuous development of genome sequencing and analysis technologies, pangenomics has gradually emerged as a new research hotspot and has been widely applied across multiple species of plants, animals, and microorganisms, providing a powerful tool for comprehensively analyzing genetic variation and diversity at the species or taxonomic group level, functional genomics, and phylogenetic reconstruction, among other studies, yielding many significant research achievements. Nevertheless, as pangenomics research is still in a developmental stage, sequencing costs and analysis expenses remain relatively high, making widespread adoption difficult; moreover, there exist unresolved issues such as inconsistent analysis standards, insufficiently comprehensive and in-depth data mining, and difficulties in applying theories to practical production, indicating substantial room for development. This article systematically summarizes the research progress of pangenomes in mining biological genetic diversity and functional genomics, mainly including their applications and research in multiple fields such as pangenome map construction, discovery of genomic variation and favorable genes, functional gene polymorphism, population genetic diversity, and phylogenetics, and explores their application potential in different domains. Simultaneously, it discusses the limitations present in current pangenome research and possible solutions, and provides prospects for its future development.

Full Text

Preamble

Application of Pan-Genome Research in Genetic Diversity and Functional Genomics

Kunli Xiang[1], Wenchuang He[1], Yi Zou[1], Dan Peng[1], Xiaoni Zhang[1], Xuezhu Liao[1], Jie Wang[1,2], Jiankang Yang[3], Zhiqiang Wu[1]*

[1]Shenzhen Agricultural Genome Research Institute, Chinese Academy of Agricultural Sciences, Shenzhen 518120, Guangdong, China
[2]School of Landscape and Architecture, Zhejiang Agriculture & Forestry University, Hangzhou 311300, China
[3]School of Basic Medical Sciences, Dali University, Dali 671000, Yunnan, China

Abstract

While a single reference genome focuses only on mining individual genetic information, pan-genome research can reflect the complete genetic information of an entire species or taxonomic group. With continuous advancements in genome sequencing and analysis technologies, pan-genomics has gradually emerged as a new research hotspot and has been widely applied across numerous plant, animal, and microbial species. It provides powerful tools for comprehensively analyzing genetic variation and diversity at the species or taxonomic level, functional genomics, and phylogenetic reconstruction, yielding many significant research achievements. Nevertheless, as pan-genomics is still in a developmental stage, high sequencing costs and analytical expenses limit its broad 普及; furthermore, issues such as inconsistent analysis standards, incomplete data mining, and difficulty in applying theoretical findings to practical production remain to be resolved, indicating substantial room for development. This review systematically summarizes research progress in pan-genome applications for mining biological genetic diversity and functional genomics, focusing on its use in pan-genome map construction, genome variation and favorable gene discovery, functional gene polymorphism, population genetic diversity, and systematic evolution, while exploring its application potential across different fields. Additionally, we discuss current limitations in pan-genome research and possible solutions, and provide perspectives on future development prospects.

Keywords: Pan-genome, structural variants, functional gene, genetic diversity, systematic evolution

---

Genetic variation is the intrinsic source of biological evolution, making the study of genetic diversity and its evolutionary patterns one of the core issues in genetics and evolutionary biology. Pan-genome research has emerged as a novel tool that comprehensively reflects species-wide genetic variation, driven by rapidly decreasing sequencing costs and rapid development of analytical technologies. Pan-genome studies enable broad exploration and utilization of genetic variation diversity at the species or taxonomic group level, representing a frontier field in modern medicine, biology, and agriculture. The pan-genome refers to the complete collection of genomic information for a species or taxonomic group, comprising two components: the core genome and the dispensable genome (also called the variable genome). Core genes are present in all individuals, while

dispensable genes exist only in some or single individuals (Figure 1 [Figure 1: see original paper]; Tettelin et al., 2005; Medini et al., 2005). The core genome, composed of sequences present in all samples, is often associated with important biological functions and phenotypic characteristics, mostly housekeeping genes that reflect species stability. The dispensable genome, composed of sequences present only in some samples, is generally associated with adaptation to specific environments or unique biological features, reflecting species diversity and specificity (Montenegro et al., 2017; Gordon et al., 2017; Wang et al., 2018; Zhao et al., 2018; Liu et al., 2020).

Currently, pan-genome research has been widely applied in numerous plant, animal, and microbial species, providing powerful tools for comprehensively analyzing genetic variation, functional gene research, and phylogenetic reconstruction at the species or taxonomic group level, yielding many significant achievements (Fu and Qin, 2012; Wang et al., 2019; Tian et al., 2019; Chen et al., 2020; Domínguez et al., 2020; Weissensteiner et al., 2020; Liu et al., 2020). However, existing pan-genomic studies have primarily focused on sequence and gene structure variation among different individuals (Montenegro et al., 2017; Zhao et al., 2018; Gao et al., 2019; Liu et al., 2020), without deeply exploring how these variations mediate changes in gene structure and function that ultimately affect phenotypes, or how these genetic differences interact with environmental factors. This review summarizes pan-genomics research progress across different species, systematically compiling its applications and studies in population genomic variation, functional gene identification and discovery, population genetic diversity, and systematic evolution, while discussing its prospects and limitations.

1. Pan-Genome Map Construction

The concept of the microbial pan-genome was first proposed in 2005 by Tettelin et al. (2005) during studies of genetic diversity in several *Streptococcus* species (GBS, group B *Streptococcus*), defining the core genome as genes present in all strains and the dispensable (variable) genome as genes present only in some strains. In GBS strains, the shared core genome accounted for 80%, with the remaining 20% representing dispensable genome information. Subsequently, Li et al. (2010) introduced the concept of the "human pan-genome" through de novo assembly and comparative genomic analysis of multiple human individual genomes, representing the sum of human population genomic information, and identified 19–40 Mb of newly discovered sequences. With the proposal and implementation of the 1000 Genomes Project, pan-genome research on human diseases has achieved many major breakthroughs, providing possibilities for precision medicine (1000 Genomes Project Consortium, 2012).

Following these pioneering studies, pan-genome map construction has been reported in numerous animal and plant species as high-quality reference genome assemblies became available. For example, a pig pan-genome map constructed from high-quality assemblies of 12 global pig breeds revealed approximately 9 Mb of pan-sequences differing between Chinese and European pig breeds, in-

cluding the essential regulator of adipocyte lipolysis TIG3 (Tazarotene-induced gene 3) (Tian et al., 2019). Pan-genome analysis of 19 wheat varieties found an average of 128,656 genes per sample, with 89,795 core genes (Montenegro et al., 2017). A tomato pan-genome map built using genomic information from 725 tomato varieties contained 40,396 genes, of which 74.2% were core genes (Gao et al., 2019). Additionally, pan-genomics has been widely applied in important plant species such as rice (Schatz et al., 2014; Yao et al., 2015; Sun et al., 2017; Wang et al., 2018; Zhao et al., 2018; Zhou et al., 2020), soybean (Li et al., 2014; Liu et al., 2020; Zhu and Huang, 2020), and maize (Hufford et al., 2012; Hirsch et al., 2014; Jian, 2017) (Table 1). Therefore, constructing species-wide pan-genome maps has become a widely applied genomic method that not only discovers comprehensive genetic information but also provides more powerful tools for functional genomics, systematic evolution, and other biological studies at the species and population levels.

2. Structural Variation and Functional Gene Discovery in Pan-Genomics

One or a few reference genomes within a species can reflect only very limited genetic variation, whereas pan-genome research can capture all variations across a species or taxonomic group, enabling studies of genome sequence and structural variation at the whole-species or taxonomic level. Genetic variation in modern biological gene pools typically includes single-nucleotide polymorphisms (SNPs), insertions and deletions (Indels), and large structural variants (SVs). SVs mainly comprise copy number variants (CNVs), presence/absence variants (PAVs), translocation events, and inversion events, which are often associated with key agronomic traits (Springer et al., 2009; Hirsch et al., 2014; Li et al., 2014; Lu et al., 2015; Zhao et al., 2018).

Comprehensive discovery of sequence and structural variations in population genomes through pan-genome analysis can identify variation sites associated with favorable phenotypes, providing important evidence for discovering and studying new functional genes. For example, a rice pan-genome constructed from 66 high-quality rice genomes identified 16,563,789 SNPs, 5,549,290 Indels, and 933,489 SVs. Analysis of genetic variation in genes related to flowering time (*Hd3a*), cold tolerance (*COLD1*), grain weight (*GW6a*), tiller angle (*TAC1*), and plant height (*Sd1*) across different materials revealed that SNP variation is the basis for variation in these genes (Zhao et al., 2018). A soybean pan-genome map constructed from 29 high-quality genomes identified 14,604,953 SNPs, 12,716,823 Indels, and 776,399 SVs (including 723,862 PAVs, 27,531 CNVs, 21,886 translocations, and 3,120 inversions), revealing that some structural variations play important roles in regulating key agronomic traits, such as PAVs, gene fusion, and Indels affecting seed coat luster, seed coat color domestication, and iron deficiency chlorosis (Liu et al., 2020).

Meanwhile, multiple sequence and structural variations discovered at different levels not only provide richer variation information but also offer more mate-

rials for studying gene functional variation. For example, through collinearity analysis among hexaploid common wheat genomes and subgenomes, researchers proposed that the "4A-5A-7B chromosome rearrangement" resulted from two chromosome translocation events and defined the precise boundaries of the rearranged genomic intervals. At the microscale, they explored the complex evolutionary history of the wheat vernalization gene *Vrn2* (*Vernalization2*), finding that the complex distribution of *Vrn2* homologous genes in the common wheat genome resulted from a series of 叠加 events including tandem duplication, polyploidization, chromosome translocation, and gene loss (Chen et al., 2020). Another study captured 238,490 SVs from 100 tomato genomes to construct a pan-structural variation (panSV) map, demonstrating that SVs underlie many transposable elements, that SV-enriched regions exhibit severe gene introgression, and that 90% of SV variations in the population can be validated in the pan-genome map (Alonge et al., 2020).

3. Functional Gene Variation and Polymorphism in Pan-Genomics

Genetic structural variation often leads to changes in gene function. Pan-genome research can comprehensively integrate relevant gene genetic information to reveal gene recombination and fusion events that result in gene function gain or loss, as well as discover new genes. For example, a candidate gene related to soybean iron deficiency chlorosis was mapped to chromosome 14. Pan-genome studies revealed two haplotypes of this candidate gene: the haplotype belonging to variety "Zhonghuang 13" is mainly distributed in low-latitude regions, while the haplotype belonging to variety "Williams 82" is mainly distributed in high-latitude regions and can survive in environments with high pH where iron exists as poorly soluble oxides. This latter haplotype has a 1.4 kb Indel in the promoter region and five variation sites in the exon region (Liu et al., 2020). In rapeseed, genome-wide PAV-GWAS (genome-wide association study) analysis revealed that PAVs in three flowering repressors—*BnaA10.FLC*, *BnaA02.FLC*, and *BnaC02.FLC*—are closely related to flowering time and ecotype differentiation. Specifically, winter rapeseed varieties all contain MITE (Miniature inverted-repeat transposable element) insertions in the *BnaA10.FLC* promoter region; 85% of spring rapeseed varieties contain LINE (Long interspersed nuclear elements) insertions in the first exon of *BnaA10.FLC*; and 81% of semi-winter rapeseed varieties contain hAT insertions in the *BnaA10.FLC* promoter region. These findings indicate that *BnaA10.FLC* determines rapeseed ecotypes and is a key gene controlling rapeseed flowering (Song et al., 2020).

Phenotypes are often the result of regulation by multiple gene networks, many of which may simultaneously influence multiple different phenotypic traits. Therefore, favorable genes for one phenotype may have detrimental effects on another. For example, the regulatory mechanisms of yield-related traits in modern tomatoes are complex. A pan-structural variation (pan-SV) study of 100 tomato genomes revealed that four structural variations led to the formation of three

MADS-Box genes that jointly affect tomato economic traits. The *j2TE* genotype exhibits a jointless pedicel phenotype that facilitates harvesting, while the *ej2w* genotype exhibits a large calyx phenotype that prevents bruising. However, the simultaneous presence of both genotypes (*j2TE ej2w*) leads to excessive inflorescence branching and low fertility. The *sb1* (*Suppressor of branching 1*) genotype can effectively overcome the negative effects of the double recessive genotype to achieve yield increases. Additionally, *sb1* genotype expression may be influenced by tandem repeats of the *STM3* gene on chromosome 1, with copy number showing dosage effects (Alonge et al., 2020). Therefore, studying the effects of gene functional variation on phenotypes in broader populations will facilitate more accurate and comprehensive evaluation of functional gene-phenotype associations, better guiding molecular breeding to develop crop varieties with stronger disease resistance, higher yield, longer shelf life, and better flavor without sacrificing other desirable phenotypic traits. Crop pan-genomics has discovered diverse correlations between numerous agronomic phenotypes and the presence, absence, and variation of specific genes (Tao et al., 2019). Research based on complete pan-genome genetic maps will help thoroughly clarify these intrinsic associations and corresponding mechanisms.

4. Applications of Pan-Genomics in Population Genetic Diversity and Phylogenetic Studies

Pan-genomics research can comprehensively analyze intraspecific genetic diversity at the genomic level, explore phylogenetic relationships among individuals and the genetic basis of phenotypic differences, and analyze genomic sequence variation and phylogenetic characteristics at the species and subspecies levels, providing evidence for important biological questions such as species origin and evolution. For example, evolutionary analysis of seven domestication-related gene loci in six rice populations using rice pan-genomes revealed that the Aus population (a subpopulation of Indica) did not all cluster on the cultivated rice evolutionary branch, leading to the proposal that the Aus rice population is in a state of incomplete domestication selection (Zhao et al., 2018). Using wheat pan-genomes to discover PAVs in 19 wheat individuals and construct phylogenetic trees revealed that the wheat variety Chinese Spring is located at the base of the evolutionary tree, providing a theoretical basis for systematic evolutionary relationships and research utilization of different wheat germplasms (Montenegro et al., 2017). A pan-genome study of 32 crow populations divided the genus *Corvus* into two major clades—Jackdaw and Crow—and explored genomic structural variation and functional traits across different evolutionary branches, particularly discovering that differences in crow feather patterns are substantial but genetic differences are minor, primarily regulated by a 2.25 kb LTR (long terminal repeats) retrotransposon insertion 20 kb upstream of the *NDP* gene (Weissensteiner et al., 2020).

Pan-genome research can also be applied to genome sequencing of germplasm resources with large differences in ecological and geographical types, mining

novel genes in species, and providing important information for candidate gene supplementation, species diversity and adaptive evolution, origin history, and invasive species studies. For example, biogeographic analysis of soybean populations revealed that modern cultivated soybeans originated in North China (Liu et al., 2020), while related studies on rice populations suggested that modern cultivated rice origins should include South China (Huang et al., 2012). Additionally, since gene banks of some crops include multiple species, particularly wild relatives with different genetic structures, constructing genetic maps containing all varieties and their relatives is necessary for broader research. Therefore, scholars have proposed the concept of the "Super-Pangenome" to explore the genetic basis and diversity of larger germplasm populations (Khan et al., 2019).

5. Future Prospects of Pan-Genomics Research

The complete genomic information of eukaryotes includes nuclear, mitochondrial, and plastid genomes. Current pan-genomics research has primarily focused on nuclear genomes, while pan-genome studies of mitochondria and plastids are gradually gaining attention. For example, researchers used whole-genome data from 2,658 cancer samples and matched normal tissue samples in the PCAWG (The Pan-Cancer Analysis of Whole Genomes) database to construct the most comprehensive mutation blueprint of the human mitochondrial genome, identifying multiple highly mutated types, among which truncated mutations were significantly enriched in kidney cancer, colorectal cancer, and thyroid cancer, suggesting that activation of special signaling pathways may have carcinogenic effects (Yuan et al., 2020). Additionally, researchers used 321 pepper chloroplast genomes to construct a chloroplast pan-genome for five cultivated species and two varieties of pepper, which not only revealed phylogenetic relationships among different *Capsicum* species through phylogenetic signal analysis but also conducted detailed analyses of genetic diversity in CDS (Coding sequence), introns, and intergenic regions of seven chloroplast pan-genomes, identifying that the intergenic region of *rpl23* and *trnI* contains a 44 bp tandem repeat and other rich variations including insertions, deletions, and single nucleotides (Magdy et al., 2019).

In some species, pan-genome research is difficult to conduct effectively due to large genome size and high proportions of mobile elements. Therefore, pan-transcriptome research focusing on all RNA information is gradually emerging. Pan-transcriptome studies have been reported in many important crops such as maize (Hansey et al., 2012; Hirsch et al., 2014; Jian, 2017) and barley (Ma et al., 2019), as well as the model organism *Arabidopsis thaliana* (Gan et al., 2011).

With the integration of multiple sequencing technologies and development of analytical strategies, pan-genomics research has experienced explosive growth. However, most studies vary in depth, and many data results still have room for further mining. Especially after constructing complete gene maps, many studies stop at identifying structural variations in a few genes without fur-

ther systematic functional research, let alone application to production practice. Additionally, with the accumulation of massive bioinformatics data, individual teams can only select partial data results for in-depth study, making it difficult to fully utilize existing data. For example, 30 years after the launch of the Human Genome Project, substantial human resources and research analysis are still needed to address more questions. Therefore, improved data sharing mechanisms and good platforms are important conditions for the healthy development and application of pan-genomics research. China has established the National Genomics Data Center (NGDC), and pan-genome data sharing platforms for some important crops or agricultural animal species have also been established, such as the pig pan-genome database PIGPAN (http://animal.nwsuaf.edu.cn/code/index.php/pan-Pig), the Chinese cabbage genome database BRAD (the Brassica database, http://brassicadb.cn), and the rapeseed pan-genome resource database (http://cbi.hzau.edu.cn/bnapus/).

Furthermore, further integration of broader, multi-level population genomic data, such as pan-genome studies across different generations and super-pangenome studies integrating multiple species, may be worthwhile new directions for exploration (Figure 2 [Figure 2: see original paper]). On the other hand, with continuous development of sequencing technologies, particularly single-cell sequencing technology, and further reduction in sequencing costs, single-cell resolution transcriptome maps have gradually begun to be applied in root development studies of rice and maize (Satterlee et al., 2020; Liu et al., 2021). Therefore, pan-genome or pan-transcriptome studies among different tissues and organs within the same individual, and even among different cells, may become new development directions (Figure 2 [Figure 2: see original paper]).

1000 GENOMES PROJECT CONSORTIUM, 2012. An integrated map of genetic variation from 1092 human genomes [J]. Nature, 491: 56–65.

ALONGE M, WANG X, BENOIT M, et al., 2020. Major impacts of widespread structural variation on gene expression and crop improvement in tomato [J]. Cell, 182: 145–1161.

BAYER PE, GOLICZ AA, TIRNAZ S, et al., 2019. Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome [J]. Plant Biotechnol J, 17: 789–800.

CHEN YM, SONG WJ, XIE XM, et al., 2020. A collinearity-incorporating homology inference strategy for connecting emerging assemblies in Triticeae tribe as a pilot practice in the plant pangenomic era [J]. Mol Plant, 13: 1694–1708.

DOMÍNGUEZ M, DUGAS E, BENCHOUAIA M, et al., 2020. The impact of transposable elements on tomato diversity [J]. Nat Comm, 11: 4058.

FU J, QIN Q, 2012. Pan-genomics analysis of 30 Escherichia coli genomes [J]. Hereditas, 34: 765–772.

GAN X, STEGLE O, BEHR J, et al., 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana [J]. Nature, 477: 419–423.

GAO L, GONDA I, SUN H, et al., 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor [J]. Nat Genet, 51: 1044–1051.

GOLICZ AA, BAYER PE, BARKER GC, et al., 2016. The pangenome of an agronomically important crop plant Brassica oleracea [J]. Nat Comm, 7: 13390.

GORDON SP, CONTRERAS-MOREIRA B, WOODS DP, et al., 2017. Extensive gene content variation in the Brachypodium distachyon pangenome correlates with population structure [J]. Nat Comm, 8: 2184.

HANSEY CN, VAILLANCOURT B, SEKHON RS, et al., 2012. Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing [J]. PLoS ONE, 7: e33071.

HIRSCH CN, FOERSTER JM, JOHNSON JM, et al., 2014. Insights into the maize pan-genome and pan-transcriptome [J]. Plant Cell, 26: 121–135.

HUANG XH, KURATA N, WEI XH, et al., 2012. A map of rice genome variation reveals the origin of cultivated rice [J]. Nature, 490: 497–501.

HÜBNER S, BERCOVICH N, TODESCO M, et al., 2019. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance [J]. Nat Plants, 5: 54–62.

HUFFORD MB, XU X, VAN HEERWAARDEN J, et al., 2012. Comparative population genomics of maize domestication and improvement [J]. Nat Genet, 44: 808–811.

HURGOBIN B, GOLICZ AA, BAYER PE, et al., 2018. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus [J]. Plant Biotechnol J, 16: 1265–1274.

JAYAKODI M, PADMARASU S, HABERER G, et al., 2020. The barley pan-genome reveals the hidden legacy of mutation breeding [J]. Nature, 588: 284–289.

JIAN YQ, 2017. Variations in pan-transcriptome and genome size in tropical Maize (Zea mays L.) and Their Applications [D]. Beijing: Chinese Academy of Agricultural Sciences.

KHAN AW, GARG V, ROORKIWAL M, et al., 2019. Super-pangenome by integrating the wild side of a species for accelerated crop improvement [J]. Trends Plant Sci, 51: 1044–1051.

LI RQ, LI YR, ZHENG HC, et al., 2020. Building the sequence map of the human pan-genome [J]. Nature Biotechnol, 28: 57–63.

LI YH, ZHOU GY, MA JX, et al., 2014. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits [J]. Nature Biotechnol, 32: 1045–1053.

LIU Q, LIANG Z, FENG D, et al., 2021. Transcriptional landscape of rice roots at the single cell resolution [J]. Mol Plant, 14:384–394.

LIU YC, DU HL, LI PC, et al., 2020. Pan-genome of wild and cultivated soybeans [J]. Cell, 182: 162–176.

LU F, ROMAY MC, GLAUBITZ JC, et al., 2015. High-resolution genetic mapping of maize pan-genome sequence anchors [J]. Nat Comm, 6: 6914.

MA YL, LIU M, STILLER J, et al., 2019. A pan-transcriptome analysis shows that disease resistance genes have undergone more selection pressure during barley domestication [J]. BMC Genomics, 20: 12.

MABIRE C, DUARTE J, DARRACQ A, et al., 2019. High throughput genotyping of structural variations in a complex plant genome using an original Affymetrix® axiom® array [J]. BMC Genomics, 20: 848.

MAGDY M, OU LJ, YU HY, et al., 2019. Pan-plastome approach empowers the assessment of genetic variation in cultivated Capsicum species [J]. Hort Res, 6: 108.

MEDINI D, DONATI C, TETTELIN H, et al., 2005. The microbial pan-genome [J]. Curr Opin Genet Dev, 15: 589–594.

MONTENEGRO JD, GOLICZ A, BAYER PE, et al., 2017. The pangenome of hexaploid bread wheat [J]. Plant J, 90: 1007–1013.

OU LJ, LI D, LV JH, et al., 2018. Pan-genome of cultivated pepper (Capsicum) and its use in gene presence-absence variation analyses [J]. New Phytol, 220: 360–363.

PINOSIO S, GIACOMELLO S, FAIVRE-RAMPANT P, et al., 2016. Characterization of the poplar pan-genome by genome-wide identification of structural variation [J]. Mol Biol Evol, 33: 2706–2719.

SATTERLEE JW, STRABLE J, SCANLON MJ, 2020. Plant stem cell organization and differentiation at single-cell resolution [J]. Proc Natl Acad Sci USA, 117: 33689–33699.

SCHATZ MC, MARON LG, STEIN JC, et al., 2014. Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica [J]. Genome Biol, 15: 506.

SONG JM, GUAN ZL, HU JL, et al., 2020. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus [J]. Nat Plants, 6: 34–45.

SPRINGER NM, YING K, FU Y, et al., 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation

(PAV) in genome content [J]. PLoS Genet, 5: e1000734.

SUN C, HU ZQ, ZHENG TQ, et al., 2017. RPAN: rice pan-genome browser for approximately 3000 rice genomes [J]. Nucl Acids Res, 45: 597–605.

TAO YF, ZHAO XR, MACE E, et al., 2019. Exploring and Exploiting Pan-genomics for Crop Improvement [J]. Mol Plant, 12: 156–169.

TETTELIN H, MASIGNANI V, CIESLEWICZ MJ, et al., 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome" [J]. Proc Natl Acad Sci USA, 102: 13950–13955.

TIAN XM, LI R, FU WW, et al., 2020. Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data [J]. Sci Chin Life Sci, 63: 750–763.

VAN DE WEYER AL, MONTEIRO F, et al., 2019. A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana [J]. Cell, 178: 1260–1272.

WALKOWIAK S, GAO L, MONAT C, et al., 2020. Multiple wheat genomes reveal global variation in modern breeding [J]. Nature, 588: 277–283.

WANG WS, MAULEON R, HU ZQ, et al., 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice [J]. Nature, 557: 43–49.

WANG YL, ZHU SS, YANG FS, et al., 2019. Pan-genome sequencing and comparative genomic analysis of atrazine-degrading bacteria [J]. Biotechnol Bull, 35: 90–99.

WEISSENSTEINER MH, BUNIKIS I, CATALÁN A, et al., 2020. Discovery and population genomics of structural variation in a songbird genus [J]. Nat Comm, 11: 3403.

YAO W, LI GW, ZHAO H, et al., 2015. Exploring the rice dispensable genome using a metagenome-like assembly strategy [J]. Genom Biol, 16: 187.

YU JY, GOLICZ AA, LU K, et al., 2019. Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars [J]. Plant Biotechnol J, 17: 881–892.

YUAN Y, JU YS, KIM Y, et al., 2020. Comprehensive molecular characterization of mitochondrial genomes in human cancers [J]. Nat Genet, 52: 342–352.

ZHAO Q, FENG Q, LU HY, et al., 2018. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice [J]. Nat Genet, 50: 278.

ZHOU Y, CHEBOTAROV D, KUDRNA D, et al., 2020. A platinum standard pan-genome resource that represents the population structure of Asian rice [J]. Sci Data, 7: 113.

ZHU GT, HUANG SW, 2020. A 360-degree scanning of population genetic variations—A pan-genome study of soybean [J]. Chin Bull Bot, 55: 56-65.

Table 1. Related researches on crop pan-genomes

Object of study	Sampling	Main research content	Reference
Rice (*Oryza sativa*, Poaceae)	3 divergent rice including Nippon-bare, IR64, DJ123	Pan-genome construction	Schatz et al., 2014
Rice (*Oryza sativa*, Poaceae)	3,010 diverse Asian cultivated rice	Pan-genome construction and structural variation	Wang et al., 2018
Rice (*Oryza* spp., Poaceae)	66 divergent rice, including cultivated rice (*O. sativa*) and wild rice (*O. rufipogon*)	Pan-genome construction, structural variation, functional gene variation and systematic evolution	Zhao et al., 2018
Rice (*Oryza sativa*, Poaceae)	12 of cultivated rice	Pan-genome construction and structural variation	Zhou et al., 2020
Maize (*Zea mays*, Poaceae)	75 wild, landrace and improved maize lines	Structural variation, functional gene variation and systematic evolution	Hufford et al., 2012
Maize (*Zea mays*, Poaceae)	503 maize inbred lines	Pan-transcriptome construction and functional gene variation	Hirsch et al., 2014
Maize (*Zea mays*, Poaceae)	31 tropical maize inbred lines	Pan-transcriptome construction and sequence (SNP) variation	Jian, 2017

Object of study	Sampling	Main research content	Reference
Maize (*Zea mays*, Poaceae)	440 inbred lines, 24 highly recombinant inbred lines and 16 F1 hybrids	Structural variation	Mabire et al., 2019
Wheat (*Triticum aestivum*, Poaceae)	Chinese wheat variety (Chinese Spring) and 18 wheat cultivars	Pan-genome construction, structural variation and systematic evolution	Montenegro et al., 2017
Wheat (*Triticum aestivum*, Poaceae)	15 *Triticum aestivum* including 10 chromosome pseudomolecule and 5 scaffold assemblies of hexaploid wheat	Pan-genome construction, structural variation and functional variation	Walkowiak et al., 2020
Barley (*Hordeum vulgare*, Poaceae)	20 varieties of barley comprising landraces, cultivars and a wild barley	Pan-genome construction and structural variation	Jayakodi et al., 2020

Object of study	Sampling	Main research content	Reference
Soybeans (*Glycine soja*, Fabaceae)	7 *G. soja* representing the geographical adaptation within the species, distributed in North, Huanghuai and South regions of China, and Japan, Korea and Russia	Pan-genome construction, structural variation, functional gene variation and systematic evolution	Li et al., 2014
Soybeans (*Glycine soja*, Fabaceae)	26 representative of soybeans, including 3 wild soybeans, 9 landraces, and 14 cultivars, and ZH 13, Williams 82 and W05 in previous studies	Pan-genome construction, structural variation, functional gene variation and systematic evolution	Liu et al., 2020
Tomato (*Solanum* spp., Solanaceae)	725 phylogenetically and geographically representative tomato, including 372 SLL, 267 SLC, 78 SP and 8 SCG	Pan-genome construction, structural variation and functional gene variation	Gao et al., 2019

Object of study	Sampling	Main research content	Reference
Tomato (*Solanum* spp., Solanaceae)	100 tomato including *S. pimpinellifolium*, *S. cheesmaniae*, *S. galapagense*, *S. lycopersicum* var. *cerasiforme* and *S. lycopersicum*	Pan-structural-variation construction and functional gene variation	Alonge et al., 2020
Pepper (*Capsicum* spp., Solanaceae)	383 cultivars, including 355 *C. annuum*, 4 *C. baccatum*, 11 *C. chinense* and 13 *C. frutescens*	Pan-genome construction, structural variation, and functional gene variation	Ou et al., 2018

Object of study	Sampling	Main research content	Reference
Pepper (*Capsicum* spp., Solanaceae)	65 samples including *C. chacoense*, *C. baccatum* var. *baccatum*, *C. baccatum* var. *pendulum*, *C. annuum* var. *annuum*, *C. annuum* var. *glabriusculum*, *C. chinense* and *C. frutescens*	Pan-plastome construction and structural variation	Magdy et al., 2019
Sunflower (*Helianthus annuus*, Asteraceae)	493 sunflower varieties including 287 cultivated lines, 17 Native American landraces and 189 wild accessions representing 11 compatible wild species	Pan-genome construction and functional gene variation	Hübner et al., 2019

Object of study	Sampling	Main research content	Reference
Cabbage (*Brassica* spp., Brassicaceae)	9 cultivated lines (*B. oleracea*) and one wild type (*B. macrocarpa*)	Pan-genome construction and systematic evolution	Golicz et al., 2016
Cabbage (*Brassica oleracea*, Brassicaceae)	Same as Golicz et al., 2016	Pan-genome construction, structural variation and functional gene variation	Bayer et al., 2019
Rapeseed (*Brassica napus*, Brassicaceae)	53 *Brassica napus* varieties including 33 nonsynthetic accessions and 20 synthetic accessions	Pan-genome construction and structural variation	Hurgobin et al., 2018
Rapeseed (*Brassica napus*, Brassicaceae)	Eight oilseed rape lines, including four SWORs (ZS11, Gangan, Zheyou7 and Shengli), two WORs (Tapidor and Quinta) and two SORs (Westar and No2127)	Pan-genome construction, structural variation and functional gene variation	Song et al., 2020

Object of study	Sampling	Main research content	Reference
Arabidopsis thaliana (Brassicaceae)	53 *Brassica napus* varieties including 33 nonsynthetic accessions and 20 synthetic accessions	Pan-NLR-gene construction, structural variation and functional gene variation	Van de Weyer et al., 2019
Poplar (*Populus* spp., Salicaceae)	3 intercrossable poplar species (*P. nigra*, *P. deltoides*, and *P. trichocarpa*)	Pan-genome construction, structural variation and functional gene variation	Pinosio et al., 2016
Sesame (*Sesamum indicum*, Pedaliaceae)	5 sesame varieties including 2 landraces (*S. indicum* cv. Baizhima and Mishuozhima) and 3 modern cultivars (*S. indicum* var. Zhongzhi13, Yuzhi11 and Swetha)	Pan-genome construction and systematic evolution	Yu et al., 2019

Note: Figure translations are in progress. See original paper for figures.