



Received: 22-08-2024

Accepted: 02-10-2024

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

Pangenome – its Aspect and Prospect

Satyesh Chandra Roy

Former Professor and Head, Department of Botany and Co-ordinator, Centre of Advanced Study for Cell and Chromosome Research and UGC Emeritus Professor, University of Calcutta, Kolkata, West Bengal, India

DOI: <https://doi.org/10.62225/2583049X.2024.4.5.3317>

Corresponding Author: Satyesh Chandra Roy

Abstract

Pangenome is a very new discipline showing a collection of unique and variable genomes of any species in one model. This discipline is a combination of three subjects of Biology like Computer Science and Applied Mathematics. All genes from all species or strains are called Pangenomes. Pangenome was first discovered during the preparation of protein-based vaccine of the G group of *Streptococcus* sp. which is the main cause of neonatal life-threatening disease by a team of scientists led by Tettelin. Conventional method of Vaccine preparation was not successful due to the existence of five different serotypes in this organism.

Then Tettelin with groups of scientists tried to use the genome sequence information for preparing protein-based vaccine in G group of *Streptococcus*. During their studies, two gap free genomes were found in this G group of bacteria. After sequencing studies of many strains, they found several regions of diversity (genomic diversity) among different isolates of same strain. After sequencing a large number of genomes, they realized that genes belong to isolates present in the species but not present in genomes. Then the Pangenome concept came to their mind and genes were classified into Core genome, Accessory or Dispensable genome and Species-specific genomes. Bacteria genomes

have a distinct number of genes but there are differences in the genomes of different strains of a single species that is mainly due to horizontal gene transfer of genes from other bacteria. The Pangenome concept and its discovery was discussed.

Plant genomes are dynamic with whole genome duplication, tandem duplication, transposable element duplication, chromosomal rearrangements or structural variations like deletion, inversion, translocation and recombination between species. Several methods have already been tried for the improvement of crop plants like Marker Assisted Selection (MAS), QTL mapping, Single Nucleotide Polymorphisms (SNP), Copy number variations and Genome Wide Association Studies (GWAS) by comparing with Reference Genome Assemblies representing the genome sequence of the genome of an individual or organism. But the Reference genome is not able to locate all genetic variations of a species. To overcome this limitation the concept of Pangenomes may help in the crop improvement which has been discussed. Pangenomic studies have already been done in many crop plants like Rice, Wheat, Maize, Brassica, Soybean etc. Some of all these aspects have been discussed.

Keywords: Pangenome, *Streptococcus*, Marker Assisted Selection (MAS), India

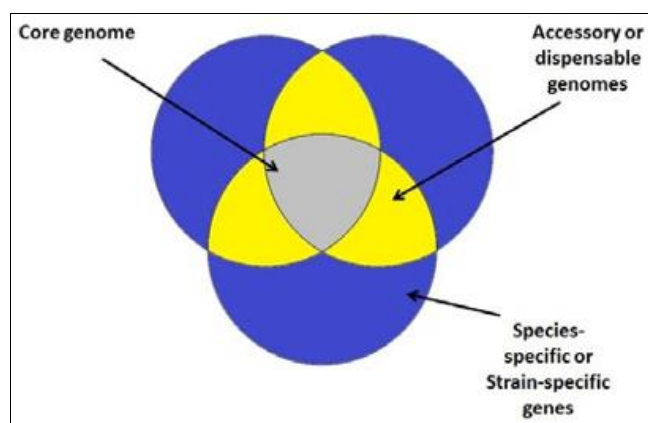
Introduction

Pangenome is a collection of the common and unique genomes of any species. It is a new discipline developed with the combination of three subjects like Biology, Computer Science and Applied Mathematics. The term Pangenome was given by Tettelin and others in 2005 to understand genome of group B *Streptococcus* (GBS) or *S. agalactiae* of different bacterial strains for the discovery of protein-based vaccine (Tettelin *et al* 2005)^[13]. This G group of *Streptococcus* is the main cause of neonatal life-threatening infections, So the genome sequence information is needed to identify proteins for the development vaccine. Tettelin and others first developed gap-free genomes in group B *Streptococcus* from the genomic diversity of 19 isolates of *Streptococcus*. (Medini *et al* 2020)^[10].

With the progress of Next Generation Sequencing, a tremendous development has been made in Whole Genome Sequencing of organisms with little time and at a reduced cost. Using this method the genome sequences of large number of organisms are stored in public databases. With the deposition of large number of genomic databases, another important discipline has been

developed which is called **Comparative Genomics**. In Comparative genomics the genomes of an organism is compared with other organisms leading mainly to establish phylogenetic relationships among species or genus. In addition to this, similarities and differences among species or different genera can be identified. When comparative genomics were used to compare the human genomes with genomes of *Drosophila melanogaster*, 548 human genes were found as homologous with *Drosophila* genes that are linked to many human diseases (cardiovascular, visual, endocrine, skeletal diseases etc.) which may open a new vistas in medical research for testing new therapies in *Drosophila* (Guimaraes *et al* 2015) [7]. In another studies it has been noted that the genetic information was variable even among closely related isolates of *Streptococcus* showing thereby that the genetic information of single isolates was not sufficient to represent the genomic content of that species (Medini *et al* 2020) [10].

Then the new idea has come to the mind of scientists to separate redundant and non-redundant genes in Genomics. Again the availability of large number of databases of whole genome sequences of prokaryotic and eukaryotic organisms lead to the idea of the concept **Pangenome**. The concept of Pangenome was first coined by Tettelin *et al* to describe the genetic constitution of taxa like species, strains, serotypes, phylum, kingdom etc (Tettelin *et al* 2005; Vernikos 2020) [13, 14]. The Next Generation Sequencing technology and innovation of different softwares in Bioinformatics including the methodology of Comparative Genomics are all responsible for the origin of new concept Pangenomics. The aim of Pangenomics is to collect the non-redundant genes of an organism. It has been noted that there are two or more genes in an organism that perform the same biochemical function and these genes are called Redundant genes.



Source: Guimardes *et al* 2015

Fig 1: Showing three parts of the Pangenome

The aim of Pangenomics is to collect Non-redundant genes of any species that can be divided into three groups such as i) **Core Genomes** where genes are shared by all species or strains of the genome and associated with fundamental aspects of an organism i.e, conserved; ii) **Accessory genomes** show variations in gene sequences where genes are present in some species but not in all species and iii) **Species /strain specific genomes** where genes are species specific (Fig 1). Genes present in accessory and species-specific zones are involved in niche adaptation (Guimaraes *et al* 2015) [7]. Niche adaptation is the process by which species can adapt or suit the environmental or ecological

changes. Pangenomic concept has made a tremendous development in bacterial genetics to understand the bacterial evolution, niche adaptation, host interaction, vaccine development, drug design and in the identification of virulence genes (Guimardes *et al* 2015) [7].

Accessory genomes can also be mentioned as **Shell genome** where genes are present in two species or individuals and the Species specific genomes are also known as **Cloud genome**.

Again Pangenomes can be divided into Open Pangenomes and Close Pangenomes. In the Open Pangenome its size is increased indefinitely when adding new genomes. In case of close pangenome the addition of new genomes does not show any increase in pangenome size.

When the pangenomes of two species are compared it will be easy to identify the unique genes and shared genes among the species as well as the phylogenetic relationship and evolutionary origin of the species.

Although the concept of pangenome was first developed in bacteria but it was also tried to develop in eukaryotic organisms like fungi, plants and animals including human. Bacterial genome is small, haploid and without introns. So the use of pangenomic concept in bacteria is not complicated. Most of the works on Pangenomics was done in bacteria. The work on bacteria will be discussed first.

Discovery of Pangenomes in Bacteria

The bacteria G group of *Streptococcus agalactiae* was causing neonatal deadly disease throughout the world. The bacteria cannot be killed by using different antibiotics. So the formation of vaccine is urgently needed. Conventional method of Vaccine preparation was not successful due to the existence 5 different disease causing serotypes in this organism. Comparative genomic studies of GBS (Group B *Streptococcus*) isolates showed genomic diversity among closely related isolates of *Streptococcus*. Vaccine of GBS is the most important as it is the cause of deadly Neonatal infections leading to death of baby. collaboration In another studies it has been noted that the genetic information was variable even among closely related isolates of *Streptococcus* showing thereby that the genetic information of single isolates was not sufficient to represent the genomic content of that species (Medini *et al* 2020) [10].

Tettelin and others (2002) [12] started research to solve the problem of this deadly disease. at the Institute for Genomic Research, USA in collaboration with Craig Rubens at Children's Hospital and Regional Medical Centre, Seattle, USA tried to use the genome sequence information for preparing protein based vaccine in G group of *Streptococcus* via maternal immunization (Medini *et al* 2020) [10]. Tettelin *et al* (2002) [12] found independently two Gap-free genomes in this G.

group of *Streptococcus* (GBS). Using microarray-based Comparative Genomic Hybridization approach the collaborative team of Tettelin identified several regions of genetic diversity among different isolates of G group of *Streptococcus* and even in the same serotype. Comparison of different strains of this G group of bacteria it was found that the core set of genes were present in all strains and also found that there are some significant number of strain-specific genes not found in any other genomes. Then the idea of Pangenomes came to the mind of scientists (Medina *et al* 2020). All genes from all species or strains are called Pangenomes.

After the discovery of Pangene in bacteria by Tettelin, a team of research with a collaboration between Ronna Rappouli's team Duccio Medini, Claudio Donati and Antonello Covacci in Italy and Herve Tettelin in USA developed protein- based vaccine using reverse vaccinology approach for this deadly disease.

After sequencing large number of genomes of GBS, it has been accepted that the genome contents can be classified into three differentiated components such as **i) Core genome** (genes present in all species), **ii) Accessory or dispensable** (genes present in several but not all) and **iii) Species specific genes** (present in one species only). This finding was found some correlation with the old observation of Geneticist Thomas Dobzhansky (Medini *et al* 2020) [10]. Dobzhansky explained how three levels could explain the processes of evolutionary genetics: 1) the origin of raw materials by mutations of genes and chromosomes, 2) the changes in populations by changes in frequencies and combination of mutations, 3) the fixation of change by reproductive isolation (Dobzhansky 1973) [4].

It has been noted that pangenome analysis can represent the total gene diversity of a population with core gene and accessory gene. The bacteria *E. coli* has been found to show maximum genetic versatility within prokaryotes. 4400 genes have been found in K12 MG 12655 strain of *E. coli* but other strains show up to additional 1000 genes encoding for different functions (Bobay 2021) [11].

Bacteria genomes have a distinct number of genes but there are differences in the genomes of different strains of a single species that is mainly due to horizontal gene transfer of genes from other bacteria. These genes can be easily identified from the analysis of pangenomes. This pangenome analysis is done from different strains of the same species of the bacteria. All coding sequences are grouped together or clustered by sequence similarity from the data of Whole Genome sequencing and then other genes are clustered into accessory groups. Generally Core genomes perform house- keeping functions while accessory genomes may involve functions of protein trafficking, defense, transposons, plasmids, phage etc. Both Core genomes and variable accessory genomes are responsible for Phenotypic traits including phenotypic variability (Innamorati *et al* 2020) [8].

Variations in the accessory genomes produce many new phenotypic characters in bacteria like diversity in capsular types, synthesis of extracellular polymeric substance (EPS), differences in Glycotransferases, EPS-biosynthesis related proteins, Methicillin resistant in *Staphylococcus aureus* giving host immunity showing thereby that accessory genomes can produce variability in bacteria leading to give adaptive potentiality in bacteria to overcome environmental stress. Similarly in *E. coli* it has been noted that variations in accessory genomes produce new ability to colonize in human cell types with genetically diverse characters helping to increase pathogenicity and adaptability in new environment. It has also been noted that gene transfer among strains of same species also can redistribute accessory genes within single strains and Horizontal gene transfer mechanism also produces diverse genotypes leading to the expansion of pangenomes (Innamorai *et al* 2020) [8].

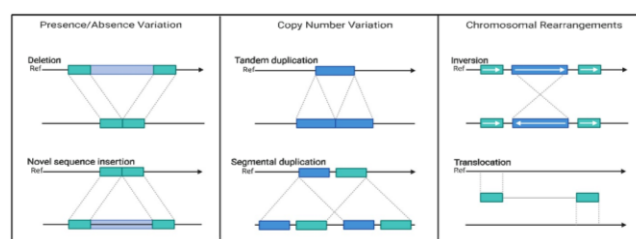
Role of Pangenomes in Crop Plants

The eukaryotic organism is usually multicellular and

contains multiple copies of same DNA (polyploids) and also with coding and non-coding genes. Structural variations are also present as has been observed mainly through single nucleotide polymorphisms (SNPs).

With the continuous global increase of human population there is a scarcity of staple food in most of the countries. As per report of FAO, about 690 million people (8.9% of the world population) are undernourished (Zanini *et al* 2021; FAO *et al* 2020) [5]. This scarcity of agricultural crop may be due to several reasons like reduce production due to climate change with extreme weather conditions, unequal distribution of crops, food waste etc.

Plant genomes are dynamic with whole genome duplication, tandem duplication, transposable element duplication, chromosomal rearrangements like deletion, inversion, translocation and recombination between species showing large number of diversity or variations (Fig 2).



Source: (Della Coletta *et al* 2021) [3]

Fig 2: Structural Variations found in Crop Genomes

These genetic and phenotypic variations mainly exhibit in agronomic traits of crop plants like phenology, yield, response to biotic and abiotic stress etc (Jayakodi *et al* 2021) [9]. The application of several modern methods has already been done in plant breeding for the improvement of crops like Marker-assisted Selection (MAS), Quantitative trait loci (QTL) methods etc. Other molecular methods have also been used to improve the production of crops like Single Nucleotide Polymorphisms (SNPs), copy number variations (CNVs), and recently genome Wide Association Studies (GWAS) where diversity sequencing databases are compared with Reference genome assemblies representing the DNA sequence of an individual's genome (Zanini *et al* 2021). Now the gene modified crops (GM crops) and gene-edited crops have already increased the production of crops in having biotic, biotic and disease resistant traits. Still it does not meet the global demand of food required for the population so there should be more technological advancement to utilise the large number of genetic variations present in the plant species. Then a new approach was done with Reference genome assemblies. Reference genome is a sequence of nucleic acid sequence database that represents set of genes standardised in one organism of a species. Reference genomes were used to identify genetic variations of a species for the development of Genome Wide Association Studies (GWAS) where diversity sequences were compared with reference genomes to identify variations associated with morphological characters. This technology has helped to form Genetically Modified Crops (GM crops).

It has been noted that the reference genome is not able to locate all genetic variations present in a species or individual (Zanini *et al* 2021). The technologies depending on a single Reference genome require modifications. So the new concept of Pangenomes was developed which is possible

due to the availability of multiple genome assemblies like microbes, plant, animal and human in public data base or repositories. Pangenomic data were divided into Core genomes (containing genes present in all species) and accessory genomes where variable genes are assembled. The analysis of pangenomes of accessory groups is used to detect agronomically related traits like yield, stress resistance genes, disease resistance genes etc leading to changes from Genome wide association studies (GWAS) to Pangenome concept.

But all these GWAS technologies are depending on a Single Reference genome which does not contain all the genetic variations present in the species or individual crop. **To overcome this limitation, the concept of Pangenome has occurred.**

Genetic information particularly the genetic variations is always needed for the improvement of any organisms including crops. Generally, genetic variation is caused by changes from Single Nucleotide Polymorphisms (SNPs) and large structural variations. In the conventional method, genetic variations studied by SNPs and Marker assisted studies (MAS) along with the data from Reference Genome are linked with phenotypic variations that can be used for the improvement of crops. But all these methods have limitations in identifying all variations including structural variations hence pangenome analysis is the only solution now. It has been noted in plants that Transposable elements (TE) are responsible for the collection of accessory genomes in Pangenomes of plant. Pangenomes in plants have already been made in Rice, Soybean, Brassica oleracea, B. napus, Capsicum sp., Medicago, Sesame, Tomato, Sunflower, Poplar, Wheat and others. The accessory genes have a role in signalling and defence systems as of disease resistance, biotic and abiotic stress etc in plants (Golicz *et al* 2020). Core genomes perform basic functions of plant like other organisms.

Pangenomes can be used here to map the accessory or dispensable or variable genomes to identify the genes for agronomic, disease resistant, stress resistant, yield etc.

It has been noted that majority of genes involved in environmental stress (biotic or abiotic) are found in the accessory or dispensable genome of Pangenomes. Thus the detection of sequences related with agronomic traits like yield, stress resistance, disease resistance, flowering time etc can be possible by using pangenomic assisted programme. On getting the sequence assemblies of whole genome of any species, pangenomic analysis is being done to differentiate the sequences into core genome and accessory or variable genomes. The core genome is present in all individuals of a species (Housekeeping Genes). Accessory genomes (Dispensable genomes) contain sequences that are present in some individuals only not in all. Core genome is for housekeeping functions like amino acid metabolism, nucleotide metabolism, lipid transport and translational machinery. Accessory genomes contain disease resistant genes, stress resistant genes and transposable elements or transposons. Both Core genomes and Accessory genomes are responsible for phenotypic traits.

Pangenomic studies on several crops like Rice, Soybean, Maize, Wheat, Brassica napus, B-oleracea etc. help to know the crop genetic diversity leading to crop improvement (Tao *et al* 2019) ^[11]. The proportion of accessory or variable genomes is found from 8% to 61% in the crops. The rice pangenome made from 68 accessions of cultivated rice

(*Oryza sativa*) and wild species (*O. rufipogon*) shows 42,580 genes with dispensable genes of about 38% (Zhao *et al* 2018) ^[16]. The dispensable genome is mainly occurred due to structural variations (SVs) of chromosomes due to variations in the sequence of DNA which is responsible for crop diversity. In addition to the occurrence of deletions, insertions, translocations, inversions of short DNA sequences. Another cause of structural variations is due to the activity of transposable elements (dynamic parts of genome). Thus the understanding of structural variations in crop species through pangenome analysis have great applications in crop improvement. Structural variations in chromosomes have an important role in the genetic control of agronomic traits (disease and stress resistance genes).

The pangenome analysis of Capsicum was made from 383 cultivars and used as a reference for PAV (Presence or absence of variations)-GWAS analysis to identify variations in carotenoid and capsaicinoid biosynthetic pathways affecting the capsaicin content in the plant. A 2.5 kb deletion of Pungent gene Pun 1 was detected in 50 cultivars with low capsaicin content. Again pangenome analysis in tomato showed the presence of shorter, non-reference allele in the Tomato LoxC promoter resulting into less desirable flavours in cultivated tomatoes which was larger in size in wild relatives (Zanini *et al* 2021).

The increase in Crop productivity depends on the study of genetic variation within the gene pool of wild relatives of crops. Pangenomic studies can easily capture all genomic content of wild relatives of crops to know and collect the genetic diversity of wild crops as Accessory or Dispensable genes in the pangenome of particular crops leading to know the full genetic diversity of that crop. The full view of structural variations in a crop from pangenomic analysis may help in understanding their impacts on gene function particularly in agronomic traits to facilitate crop improvement (Tao *et al* 2019) ^[11].

During Crop improvement studies it has been noted the importance of Transposable elements as it is the driver of producing structural variations in chromosomes. Transposable elements can modify the structure and amount of gene product after transcription. So the insertion of transposable element in the gene can either silence the gene or enhance the gene product. Figure can be given see Plant Pangenome 3) Thus transposons may play an important role in crop improvement. For example, Harbinger-like transposon in maize promotes flowering by repressing the expression of ZmCCT9 gene. Actually Harbinger DNA transposon encodes a DDE transposase and a SANT/Myb/trihelix domain containing a DNA binding protein. There are many types of this domain that helps in switching defective protein, nuclear receptor co-repressor, transcription factor that couples histone tail binding to catalytic activity through chromatin remodelling. In Rice Gypsy retrotransposon has been found to enhance the expression of OsFRDL4 gene to promote aluminium tolerance. In Tomato also Copia retrotransposon, Rider, has brought polymorphisms in the SUN locus forming the oval shape of the Roma tomato variety (Delta Colleta *et al* 2021) ^[3]. All these changes in agronomic traits of crop plants are the effect of transposons through insertion and deletion of genes in crop genomes resulting in the changes of gene product. Previously all researches in this line are done through QTL mapping using reference genome. But as the reference genome cannot detect all variants associated with

traits so it is not possible to get a total picture of genomes through QTL mapping. In Pangenome model, all genes of variants (generally found in wild crops) of a genome are collected in the section of Accessory genes (Dispensable genes) facilitating the use of variants in plant breeding for crop improvement.

Thus the use of Pangenome graph or model as a reference may help in getting greater integration and varying data types for functional and comparative analysis to develop high performance crops (Zanini *et al* 2022)^[15].

However it is noteworthy that eukaryotic organisms are very complex containing some types of RNA encode some genes or traits. These RNAs are tRNA, rRNA, snoRNA, scRNA, microRNA, and siRNA. Although we are already getting benefits in Crop Plants using Pangenome analysis, but the complete Pangenomic model including the study of DNA and RNA sequences may definitely show more potentialities not only in different disciplines of plant sciences but also in animals and human being.

References

1. Bobay Louis-Marie. The Prokaryotic Species Concept and Challenges. pp. 21-46. In the Book: The Pangenome Diversity, Dynamics and Evolution of Genomes. Eds. Herve Tettelin and Duccio Medini. Springer Open, 2020. ISBN 978-3-030-38281-0.
2. Bosi E. *Et al*. Defining orthologs and pangenome size metrics. *Methods in Mol. Biol.* Pages missing, 2015.
3. Della Coletta Rafel, Yinjie Qiu, Shujun Qu, Matthew B Hufford, Candice N Hirsch. How the Pangenome is changing Crop genomics and improvement. *Genome Biology*. 2021; 22:3. Doi: <https://doi.org/10.1186/s13059-020-02224-8>
4. Dobzhansky T. Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teach.* 1973; 35:125-129.
5. FAO, IFAD, UNICEF, WFP, WHO. State of food security and nutrition in the world. 2020: Transforming food systems for affordable healthy diets. Food and Agriculture Organization, 2020. Doi: <https://doi.org/10.4060/ca9692en>
6. Golicz Agnieszka A, Philipp E Bayer, Prem I Bhalla, Jacqueline Batley, David Edwards. Pangenomics comes of age—from bacteria to plant and animal applications. *Review*. 36(Issue 2):132-145. Doi: <https://doi.org/10.1016/j.tig.2019.11.006>
7. Guimardes Luis Carlos, Laendro Benevides de Jesus, Marcus Vinicius Canario Viana *et al*. Inside the Pan-Genome – Methods and Software Overview. *Current Genomics*. 2015; 16:245-252.
8. Innamorati Katherine A, Joshua P, Earl D, Aggarwal Garth D, Ehrlitch N, Luisa Hiller. The Bacterial Guide to Designing a Diversified Gene Portfolio. Pp. 51-87. In the Book: The Pangenome Diversity, Dynamics and Evolution of Genomes. Eds. Herve Tettelin and Duccio Medini. Springer Open, 2020. ISBN 978-3-030-38281-0.
9. Jayakodi, Murukarthik, Mona Schreiber, Nils Stein, Martin Mascher. Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Research*. 2021; 28(1):1-9.
10. Medini D, Claudio Donati, Rino Rappuoli, Tettelin H. The Pangenome: A Data- Driven Discovery in Biology. In the Book: The Pangenome Diversity, Dynamics and Evolution of Genomes. Eds. Herve Tettelin and Duccio Medini. Springer Open, 2020. ISBN 978-3-030-38281-0.
11. Tao Yongfu, Xianrong Zhao, Emma Mace, Robert Henry, David Jordan. Exploring and Exploiting Pangenomics for crop improvement. *Mol.Plant.* 2019; 12:156-169.
12. Tettelin H, *et al*. Complete genome sequence and comparative genomic analysis of an emerging human pathogen. *Proc. Natl. Acad. Sci. USA*. 2002; 99:12391-12396.
13. Tettelin H, Massignani V, Cieslewicz MJ, Donati CD, Medini NL Ward SV, *et al*. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae* implications for the microbial “pan-genome”. *Proc. Natl.Acad.Sci. USA*. 2005; 102:13950-13955.
14. Vernikos GS. A Review of Pangenome Tools and Recent Studies. In the Book: The Pangenome Diversity, Dynamics and Evolution of Genomes. Eds. Herve Tettelin and Duccio Medini. Springer Open, 2020. ISBN 978-3-030-38281-0.
15. Zanini, Silvia F, Phillip E, Bayer, Rachel Wells, Rod J, Snowdon, *et al*. Pangenomics in crop improvement----- from codinf structural variations to finding regulatory variants with pangenomr graphs. *Plant Genome*. 2022; 15:e20177. Doi: <https://doi.org/10.1002/tpg.2.20177>
16. Zhao Q, Feng Q, Lu HY, Li Y, Wang A, *et al*. Pangenome analysis highlights the extent of genetic variation in cultivated and wild rice. *Nat. Genet.* 2018; 50:278-284.