Pangenomics & Its Role in Crop Improvement

Bhagyasri Majhi[1], Yogesh Kashyap[1], Manisha Kumari[2] and Boddu Sangavi[1]

[1] Ph.D. Research Scholar, Department of Genetics and Plant Breeding, N. M. College of Agriculture, Navsari Agricultural University, Navsari, Gujarat-396450.

[2] Ph.D. Research Scholar, Department of Genetics and Plant Breeding, Sardarkrushinagar Dantiwada Agricultural University, Gujarat-396450.

Corresponding Author: majhibhagyasri2@gmail.com

Pangenomics, the study of the entire genomic diversity of a species or clade, has revolutionized the field of crop improvement. By analysing the pangenome of crops, researchers can identify novel genes, alleles, and genomic variations that contribute to desirable traits such as yield, disease resistance, and climate resilience. This information can be leveraged to develop new crop varieties with improved characteristics, enabling farmers to increase productivity and adapt to changing environmental conditions. Pangenomics also facilitates the discovery of hidden genetic diversity within crop species, allowing breeders to tap into this diversity to develop more resilient and sustainable crops. Furthermore, pangenomic approaches can aid in the identification of genetic markers associated with specific traits, enabling marker-assisted breeding and accelerating the crop improvement process. Overall, pangenomics holds great promise for transforming crop improvement, enabling the development of crops that can meet the food security challenges of the future.

Pangenome

Pangenome = Pan + Genome, “Pan" meaning "whole “Research findings, term ‘pan-genome’ was first used by Sigaux in 2000 while conceptual outline was given by Dr. Herve Tettelin in 2005.

A pangenome is the entire set of genes and genomic sequences present in a species or a group of related species. It includes the "core" genes that are shared among all individuals of a species, as well as the "dispensable" or "variable" genes that are present in some, but not all, individuals. The pangenome concept was first introduced in 2005 by researchers studying the genomic diversity of bacteria. Since then, the concept has been applied to various organisms, including plants, animals, and fungi.

The pangenome typically includes:

1. **Core genome:** The set of genes that are present in all individuals of a species and are essential for the species' survival and function.
2. **Dispensable genome:** The set of genes that are present in some, but not all, individuals of a species. These genes may be involved in adaptation to specific environments or lifestyles.
3. **Variable genome:** The set of genes that are present in different versions or alleles in different individuals of a species.

Genome V/S Pangenome

Genome

A genome is the complete set of genetic instructions encoded in an organism's DNA. A genome represents a single individual or a specific strain of an organism. A genome includes all the genes, non-coding regions, and repetitive elements present in an individual's DNA. Genome size varies greatly among organisms, ranging from a few million base pairs in bacteria to billions of base pairs in plants and animals.

Pangenome

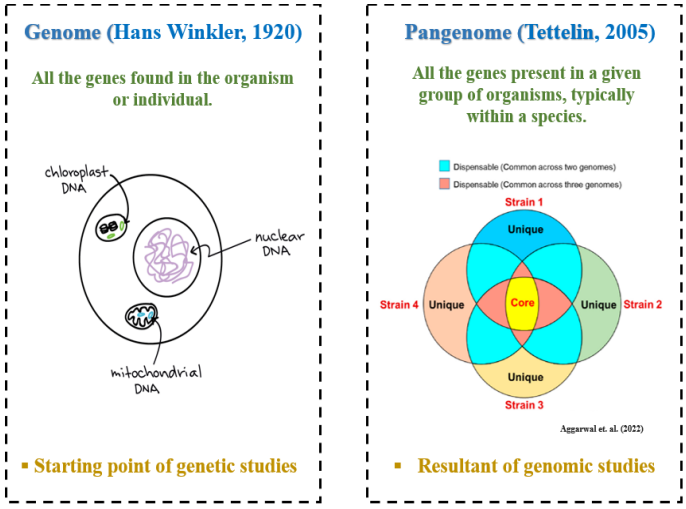


Fig 1: Representation of genome & pangenome (Bayer *et al.* 2019)

A pangenome is the entire set of genes and genomic sequences present in a species or a group of related species. A pangenome represents a collection of genomes from multiple individuals or strains within a species or clade. A pangenome includes the core genome (shared among all individuals) and the

dispensable genome (variable genes and sequences present in some but not all individuals). Pangenome size is typically larger than a single genome, as it encompasses the cumulative genetic diversity of a species or clade.

Types of Pangenome

Here are the main types of pangenomes:

1. Closed Pangenome

- ✓ Pangenome size increases with the addition of new genomes.
- ✓ New genes are frequently discovered, contributing to the expansion of the pangenome.
- ✓ Genetic diversity within the population or species is high.
- ✓ Calls for more sequencing
- ✓ Core genome-small
- ✓ Eg. Plants

2. Open Pangenome

- ✓ After the inclusion of a sufficient number of samples, size is fixed and doesn't change with the addition of new genomes.
- ✓ No new genes are discovered upon sequencing the additional genome.
- ✓ No need for more sequencing
- ✓ Core genome-large

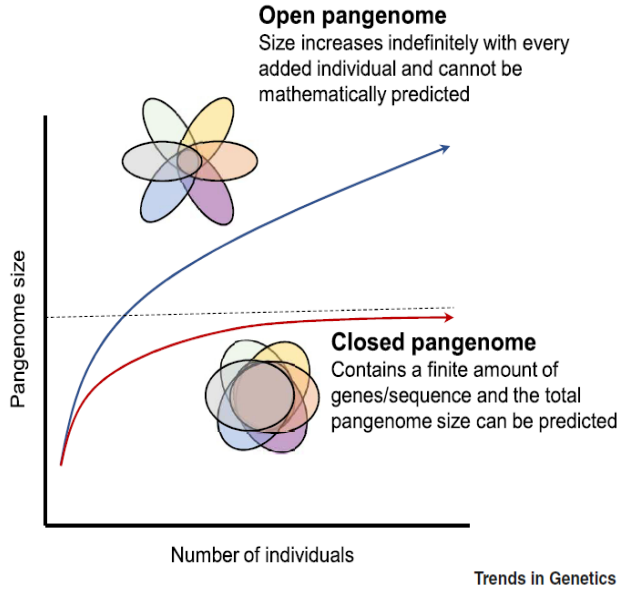


Fig 2: Difference between genome and pangenome (Golicz et. al., 2020)

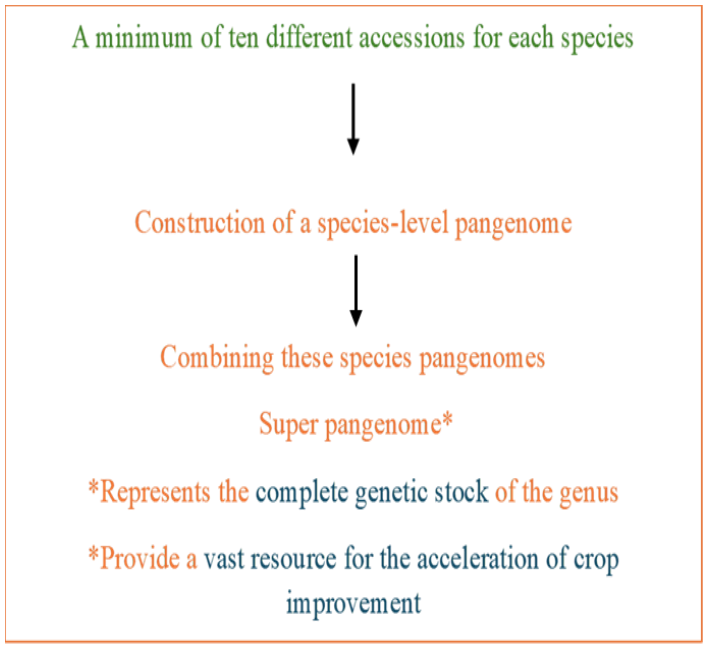
3. Super-Pangenome

- ✓ **Extension of the traditional pangenome concept.**
- ✓ Entire set of genes in a given species.
- ✓ Reflect the whole genomic architecture of that genus.
- ✓ Valuable tool for **evolutionary research.**
- ✓ Broader representation of genetic diversity compared to a single-species pangenome.
- ✓ Useful to **transfer genes from** the species belonging to **distantly related gene pools.**

Scheme for construction of super pangenome

1. **Core genome construction:** Construct the core genome by identifying genes present in all strains or individuals.
2. **Pan-genome construction:** Construct the pan-genome by combining the core genome with the dispensable genome (genes present in some, but not all, strains or individuals).
3. **Super pangenome construction:** Construct the super pangenome by integrating the pan-genomes of multiple species or clades.

Utilization of pangenome in crop improvement



1. Pan genomics in utilizing crop wild relatives

Pan genomic studies characterize full genomic content of CWRs uncovering genetic diversity including SVs and SNPs, that defines uniqueness of individual CWRs laying the foundation for utilization of this untapped resource.

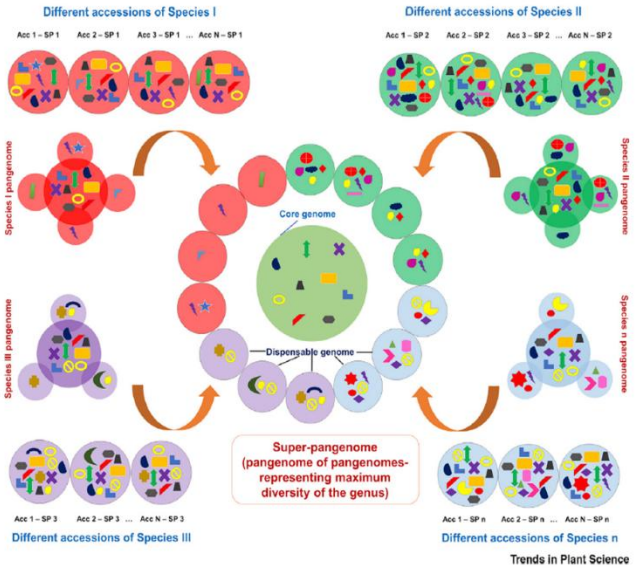


Fig 3: Construction of a super-pangenome using diverse accessions of all of the species of a given genus (Khan et. al., 2020)

2. Domestication of new and existing species

- Utilization of pan genome information in identifying causal variants underlying domestication traits. Variation for fruit weight QTL fiw3.2 caused by tandem duplication of the cytochrome P, gene elucidated with the help of tomato pangenome.
- De novo domestication of wild plants to utilize genetic variation from secondary and tertiary gene pool. Zsogon *et al*. edited six loci [*SELFPRUNING, OVATE, FASCIATED, FRUIT WEIGHT 22, MULTIFLORA, LYCOPENE BETA CYCLASE*] in Solanum pimpinellifolium significantly increased its yield, productivity, and nutritional value

3. Easing Mapping and GWAS

- A major concern in QTL mapping and GWAS based on SNPs from a single reference genome is reference bias,

A maize gene conferring resistance to sugarcane mosaic virus is identificd by GWAS using markers based on B73 but not present in PH207 genome assembly.

- In pangenome perspective, the contribution of SV's te trait variation is clear. In

Brassica napus, GWAS was performed with PAVs identified from eight whole-genome assemblies, and causal associations between SVs and silique length, seed weight, and flowering time were discovered that were not captured by SNP-GWAS.

- GWAS conducted using rice pan-genome genotyping array (RPGA) includes 80K genome-wide SNPs genotyping data of a rice diversity panel detected 1otal of 42 loci, regulating grain size/weight traits in rice.

4. Using pan genomics to dissect agronomic traits

* Dispensable genome is enriched with genes involved in environmental responses.

* Pangenomes are being increasingly used in the detection of sequences associated with agronomically relevant traits, such as yield and stress resistance.

* This lead to the transition from so-called genomic-assisted to pan genomic- assisted breeding strategies like:

1. Resistance to biotic stress/Disease resistance
2. Vernalization and flowering time
3. Fruit, grain, yield and seed quality
4. Abiotic stress tolerance and Resistance to abiotic stress
5. Plant architecture

5. Complexity of polyploid genomes

- High quality genome assembly of polyploid species has been difficult to achieve due to their inclusion of multiple, closely related sub genomes and the associated challenges in discriminating homeologous loci and creating non-mosaic sub genome scaffolds.
- Many have resorted to sequencing diploid progenitors or closely related species.
- However, closely related diploids fail to capture lineage specific SNPs, SVs, and other forms of variation that have accumulated post polyploidization.
- The degree of structural variation within polyploid species is greater than diploid species, and SVs may be particularly fruitful markers for genomic approaches to polyploid crop improvement.

6. Editing regulatory elements for crop improvement

- CREs constitute attractive genome editing targets.
- In tomato, inspired by the natural variation observed between wild and domesticated

relatives, researchers used CRISPR-Cas9 mutagenesis to generate novel alleles of the promoter of SICLV3, a gene affecting fruit size and engineered a continum of phenotypic variation.

- In maize, it was possible to engineer a spectrum of variation for yield-related traits by targeting homologues of CLV3

Future Prospects

- The pan genomic studies can provide plant geneticists and breeders a comprehensive genome resource.
- New tools are required to support variation graph assembly, pangenome construction and visualization.
- An integrated pangenome browser should be developed, capable of representing SNPs and SVs in multi reference coordinate system for genome analysis.
- Expanding the pangenome beyond species will increase the use of wild gene sequence diversity in crop improvement.

Conclusion

- The dispensable genome (Structural variants) is not essential for survival.
- Pangenomes can be used to map the dispensable genome, highlighting which varieties host genes that provide an agronomic benefit.
- The application of pangenome wide SNPs as well as the inclusion of PAV in the association analysis, increases the power of genomic associations and the identification of causal variants for agronomic traits.
- The variants identified from pan-genomes can support genome editing approaches.
- Adoption of pangenome graph as a reference facilitates comparative analyses to develop next generation of climate resilient and high-performance crops.
- Once pangenomes are available for large number of diverse species, we can understand how species and higher taxa are defined at genome level, providing insights into plant evolution and domestication.

References

Bayer, P.E., Golicz, A.A., Tirmaz, S., Chan, C.K.K., Edwards, D and Batley, J. 2019. Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome. *Plant Biotechnology Journal*, **17**(4):789-800.

Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J and Edwards D. 2020. Pan genomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends in Genetics*, 1-14.

Khan, A. W., Garg, V., Roorkiwal, M., Golicz, A. A., Edwards, D & Varshney, R. K. 2020. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends in plant science*, **25**(2): 148-158.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L & Fraser, C. M. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, **102**(39):13950- 13955.

* * * * * * * * *