Plant Pan-genomes: A New Frontier in Understanding Genomic Diversity in Plants

ABSTRACT

The comparison of several associated species and plant genome sequencing efforts has increased in recent years. The inflated level of the genomic variety leads to the discovery that the single reference genomes may not reflect the variability in a species, resulting in the evolution of a pan-genome idea. Pan-genomes exhibit a species' genetic variability and contain mutant genes lacking in some individuals and essential genes present in all individuals. Mutant gene classifications often reveal cross-species parallels, including genes for abiotic and biotic stresses generally concentrated within mutant gene groupings. Here we discuss the history of pan-genomics in plants, investigate the causes of gene variation, deletion, and existence and demonstrate why pan-genomes might assist crop genetics and breeding research.

Keywords: Pan-genome; presence and absence variations; whole-genome assembly; mutant gene; polyploid; transposable elements; homoeologous exchange;

1. INTRODUCTION

The notion of pan-genomes was initially introduced in bacteria in 2005 [1] when the sequencing of numerous strains of *Streptococcus agalactiae* identified an important genome characterized by 80% of *S. agalactiae* genes, while the remaining 20% missing in at least one isolate [1]. However, it took a decade following the initial bacterial pan-genome effort to develop plant pan-genomes. Although this was mainly owing to the significant costs associated with collecting and analyzing data, it was also anticipated that there would be very little gene presence and absence variations (PAVs) in higher species since they transfer genetic material not as easily as microbes [2]. When it was first utilized in plants in 2007, it was used to find small mutant sections of the rice and maize genomes that differed from each other [3]. However, because of the absence of precise whole-genome assemblies for numerous individuals of the same species at the time, it was impossible to determine the extent to which genes were present or absent. When the cost of DNA sequencing decreased, it became practical to compare complete genomes of different individuals, and three broad methodologies for supra-genome assembly were devised [3] (Figure 1). The initial approach was whole-genome assembly (WGA) and comparison, which involves assembling and comparing the genomes of various individuals [4,5]. This was then supplemented by the repeated assembly and, PAV strategy, in which genomic sequences from several species are matched to a database, and unaligned data is assembled and contributed to the increasing pan-genome reference. Following that, entire readings are remapped to the pan-genome, which allows for accurate PAV calling throughout the population. In recent years, a lot of progress has been observed in graph-regulated pan-genome assembly, which constructs a network expressing genetic conservation and variability [6].

In addition to being highly complementary, the WGA and comparison strategy provides critical structural and gene whereabouts. In contrast, the recurring assembly technique allows the extension of research to exceptionally multiple participants, identifying infrequent genes and the dispersion of PAV in a population.

It has only recently become possible to use graph assembly in more complicated genomes. If we study microbes, graph assembly has been used a lot since long-read DNA sequencing has become more credible.

Each of these methods has advantages and disadvantages; for example, a repetitive assembly cannot distinguish between high sequence variation at a locus and sequence addition or deletion, whereas WGA cannot distinguish among true genome diversity across individuals and frequent mistakes and variation reported in assembling and annotating methods. The graph-based analyses of plant genomes are just valuable for a small number of plant genomes right now because they require a lot of memory and data space, and they also have other limitations.

Recently, a lot of focus on plant supra-genome investigations is observed (Table 1). In order to create the first plant pan-genome, seven wild soybean individuals were contrasted for the first time [7]. The results revealed a significant variation in seed yield and size, seed structure, days to flower and maturation time, and further copies of disease tolerance genes in the wild *G. soja*. When 18 *Arabidopsis thaliana* accessions were evaluated [8] in a prior study, the researchers focused on protein isoforms and gene expression instead of gene availability and unavailability. In parallel with the soybean genome assemblies, a loss in the S5 hybrid infertility region in a single genotype and PAVs in the subsurface resistance gene Sub1A were detected in a short rice supra-genome derived from four different genotypes [9].

In a recent study of seven de novo constructed *Brassica napus* genomes, two PAVs, each representing hAT retrotransposon incisions inside known flowering time genes, were discovered to be linked with flowering time [10]. Another research in *A. thaliana* found non-syntenic hotspots of rearrangements (HOTs) linked to sequential duplications based on seven assemblages [11]. There is less meiotic recombination in these HOTs, and they have fewer genes, with disease resistance genes being substantially more prevalent. Mining for HOTs in other plant species is presently not viable owing to a lack of high genomic assembly for several individuals of a species. There has not been any evidence of a link between disease tolerance and hotspots for rearrangement.

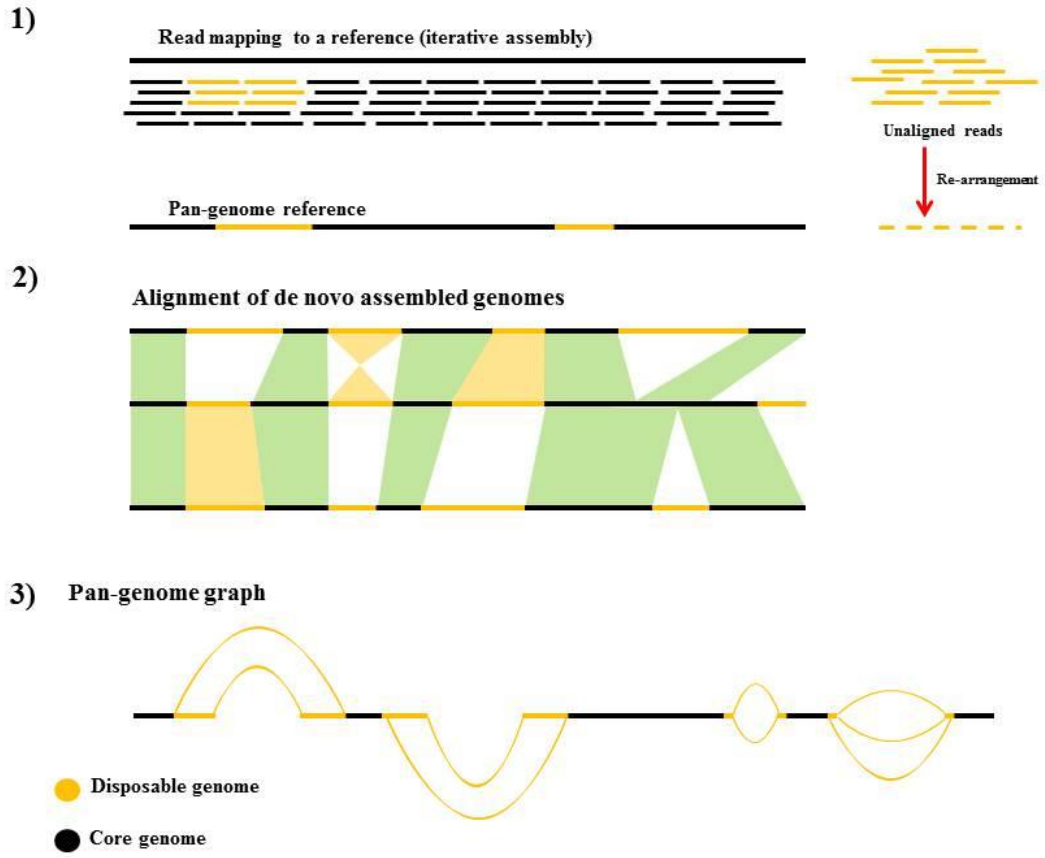

Fig. 1. Comparison of Pan-genome approaches

A number of pan-genome studies, including *Brassica oleracea* (based on 10 individuals [12]), bread wheat (based on 18 individuals [13]), and canola (based on 53 individuals [14]) have used the iterative mapping and assembly approach [4] to reduce the expense of producing high-quality assemblies for numerous individuals. Two important findings were made in these initial plant pan-genome studies: each species studied had between 15% and 40% of its total number of gene content, and genes that make PAV are often linked to biological and environmental stress resistance.

Other recent research has included the analysis of a pan-genome using 54 *Brachypodium distachyon* plants, which discovered additional 7,134 genes and found that certain mutant genes are core genes inside sub-populations [15], hence sustaining the population structure. Five sesame plants were utilized to generate a sesame pan-genome [16], allowing for genetic differentiation among traditional and contemporary sesame varieties. This suggests that pan-genomes may be used to identify alterations in the gene frequencies throughout cultivation and breeding. There were three genes associated with seed weight in the latest pan-genome research of 90 individuals of *Cajanus cajan*, demonstrating that presence and absence variations (PAVs) may be used to enhance SNPs for trait correlation [17].

There is a rising interest in the dispersion of mutant genes in populations as a result of the growing popularity of pan-genome studies. More than 10,000 new genes were identified in a rice supra-genome study of 66 representative samples from 387 wild *Oryza rufipogon* and 1100 *O. sativa* genotypes [9]. Additionally, they confirmed prior findings from three rice accessions [18] linking submergence tolerance and phosphate deficiency tolerance genes. Using 725 distinct tomato lines, a pan-genome analysis was conducted [19] that discovered 4,873 genes, mostly were associated with disease tolerance. The study also discovered an atypical allele associated with *Solanum lycopersicum* flavor that was chosen during domestication but re-emerged in current *Solanum lycopersicum* varieties as a result of wild introgressions. There has recently been an investigation into the diversity associated with agronomic traits in soybean using a pan-genome research that united the assembly of 27 lines with re-sequencing data from 2,898 different lines [20].

2. PAN-GENOMES AND PLANT BIOLOGY

Many of the plants used for reference genome sequencing have historical significance; such as the *Triticum aestivum* cultivar Chinese Spring was chosen for reference genome sequencing because it was this cultivar that the current karyotype system was developed [21]. The genes of Chinese Spring, on the other hand, are considerably different from that of current cultivars. Using the initial *T. aestivum* pan-genome research, scientists discovered 11,840 genes, found in 22 re-sequenced current types but missing from Chinese Spring [22]. The usage of a single reference (SR) may influence our knowledge of the genetic foundations of phenotypes; such as, the Lr49 *T. aestivum* rust tolerance gene exhibits unusual structural diversity across cultivars [19]. A wide range of genomic studies will get better if they use pan-genomes as references. For instance, utilizing a pan-genome assembly increases short-read mapping efficiency over an SR, leading to better mutant calls and more exact measurements of gene expression [23-25]. It is still difficult to distinguish plant species solely on gene encoding, especially when there is a lot of variation in gene PAV across individuals. However, when more species' pan-genomes are created, a better knowledge of gene retention and deletion might well aid in establishing species-level modifications in gene expression.

Gene PAV can be used to boost agricultural yields and plays an important role in the study of fundamental biology. More than 30% of the advances in crops production in the late 20th century were attributable to crop wild relatives utilization in the crop breeding programs [26]. Crop wild relatives frequently have a greater collection of genes and represent a rich foundation of genetic diversity for crop breeding. With the use of pan genomic research, we can examine gene retention and deletion during breeding and adaptation [27], which assists in discovering dispersed variability and the capability of incorporating genes into current cultivars. Such as, gene deletion related to flavor, which occurred during the introduction of tomatoes in Chile, Mexico, and Brazil, has recently been incorporated into new varieties [28]. Gene distribution research among wild species in various settings might aid in the development of agricultural plants that are more adaptable to a variety of environments and more resilient to climate change [29].

Table 1. Description of various plant pan genome studies

Species	Approach	Domestication status	Ploidy	Number of accessions	Pangenome genes	References
Soybean	De novo	Crop	Diploid	204	3,621	[10]
Maize	Iterative	Crop	Diploid	503	8,681	[11]
Maize	De novo	Crop	Diploid	2	-	[12]
Wheat	Iterative	Crop	Hexaploid	18	140, 500	[13]
Soybean	De novo	Wild and crop	Diploid	26	57,492	[14]
Brassica rapa	De novo	Crop	Diploid	3	41,858	[15]
Glycine soja	De novo	Wild	Tetraploid	7	59,080	[16]
Populus	Read mapping	Wild	Diploid	7	-	[17]
Brassica napus	Iterative assembly	Crop	Tetraploid	53	94,013	[18]
Pepper	Iterative assembly	Crop	Diploid	383	51,757	[19]
Tomato	Iterative assembly	Crop	Diploid	725	40,369	[20]
Juglans (walnut)	De novo	Wild	Diploid	6	26,458	[21]

Disease resistance genes have a negative impact on fitness [30]. However, recent discoveries of multiple genes across plant supra-genomes have revealed that they are accumulated for genes engaged in adaptations to abiotic and biotic stress, specifically for genes associated with disease resistance. Wheat [31], rapeseed [32], wild cabbage [33], and tomatoes [34] are all examples of monocots and dicots that have been shown to contain a wide range of disease-resistant genes and human-based pan-genomes have also reported similar results [35,36]. The discovery of the NLR disease resistance genes has given birth to the idea of a pan-genome research study called the pan-NLRome [37]. It has only been used in *A. thaliana* [38] until now, and only 37 out of 64 accessions were adequate to get 90% of the NLR genes. A large number of disease resistance genes are clustered together in physical clusters [39–42], with some of them exhibiting significant variation [43,44]. Clustered genes may be different from unclustered genes because of the irregular crossing over and meiotic disruption due to orthologous repeats in these clusters [45]. This happens only for specific kinds of disease tolerance genes (type I), whereas type II genes prove only a few genomic changes in wild cabbage [46], which is similar to findings in *Arabidopsis thaliana* [47].

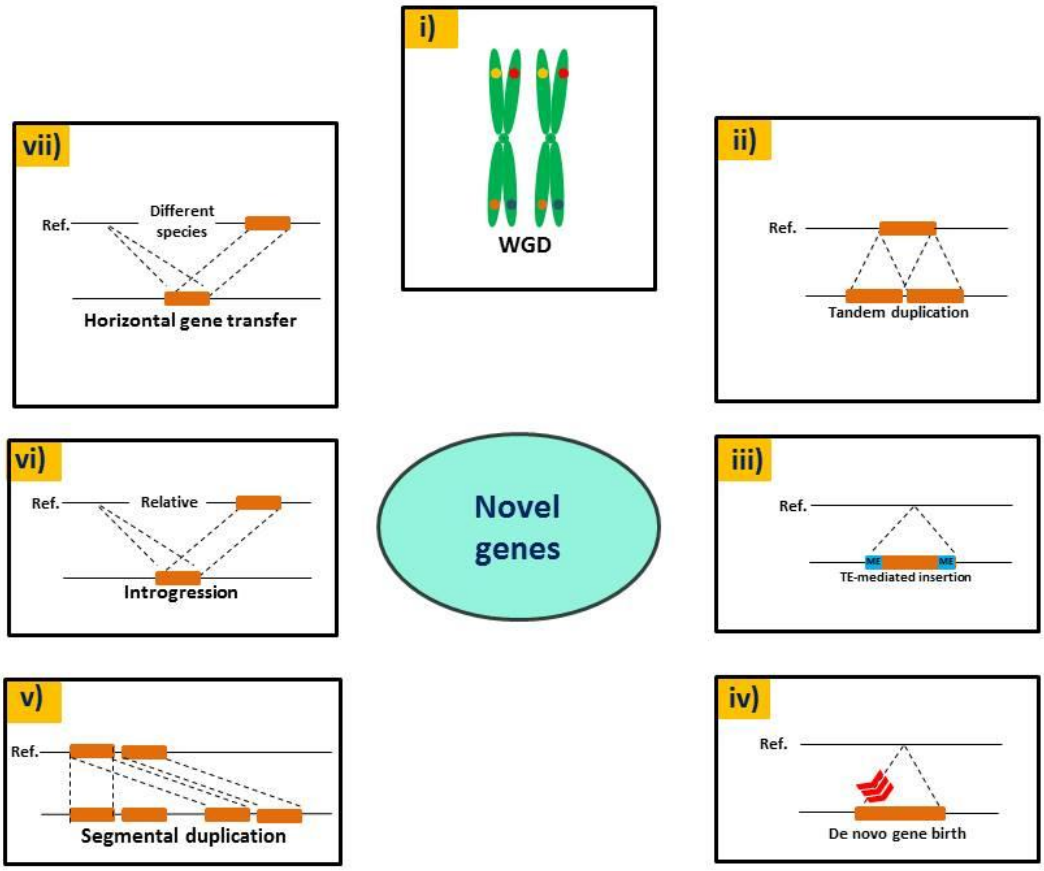

Fig. 2. Different sources for novel genes

While disease resistance genes are abundant in the mutant gene fraction in many plant species, this is not true in all. The *Amborella* pan-genome, for example, has a small number of disease resistance genes, most of which are core genes, which might represent the species' unique geographic location and evolutionary record. Abiotic stress and environmental adaptability are frequently connected with mutant genes [19,32], implying that these genes may be useful in crop breeding techniques in the future.

3. INCEPTION OF MUTANT GENES IN PLANTS

Even though the predominance of the gene PAV has been well documented, the genesis of mutant genes is still a subject of much debate. Several processes of gene losses and gains have been identified in plants, all of which have the potential to contribute to mutant gene creation (Figure 2). Introduction from similar species, horizontal gene transmission, and de novo gene formation are all ways to get new genes [48-52]. As a consequence of deletions, such as those resulting from intra-chromosomal recombination and pseudogenization, genes may potentially be lost [12,53,54]. The study of pan-genomes in plants gives a complete picture of how different mechanisms of addition and deletion affect the total number of genes in organisms, as well as how adaption can change the frequency of different genes.

There appear to be a higher fraction of mutant genes in polyploid species compared to diploid species; however, there are presently not many polyploid pan-genomes accessible to corroborate this tendency. Dominant sub-genomes can affect allopolyploid gene content, as seen in *Fragaria* × *ananassa* [55,56] and *B. napus* [57,58], with the dominant sub-genomes containing a higher number of significant genes. In the case of WGDs (whole-genome duplications), the entire gene complement is doubled and is often followed by gene loss, renowned as fractionation. After differential fractionation, the *B. lineage* had a whole-genome triplication, which produced three sub-genomes: LF; the lowest fractionated, MF1; the most fractionated first, and MF2; the most fractionated second [59,60]. The pan-genome of *B. oleracea* showed that there was a statistically significant relationship between the sub-genome assembly and the number of mutant genes, with MF2 having the most and LF having the fewest [61]. On extremely short evolutionary time scales, the sub-genomic position of mutant genes in canola is expected to correlate with the rate of gene loss, which connects with an intraspecies variation.

Moreover, most of the mutant genes in *B. oleracea* were not allocated to sub-genomes [62,63], echoing a finding in *Brachypodium* that mutant genes are less syntenic with orthologous sections in other grasses, showing that they are developed outside of syntenic blocks [64]. *Sesamum indicum*, which experienced whole-genome duplication roughly 80 million years ago, was the subject of pan-genome research that sought to determine the origin of core and mutant genes. Whole-genome duplication was shown to be responsible for more than one-third of the core genes and just around 12% of the mutant genes. Many of the mutant genes assigned to the whole-genome duplication origin, do not occur in syntenic blocks, which explains the low fraction of mutant genes attributed to the whole-genome duplication origin. Local tandem duplications could be attributed to a comparable percentage of the core and mutant genes (both 10 percent), indicating that, for *S. indicum*, tandem duplications (TD) are not a substantial cause of mutant genes, despite the presence of line-specific variations [65].

Another typical source of gene PAV in amphipolyploid plants is the homoeologous exchange (HE), which occurs in amphidiploid plants [23,66]. As previously mentioned, rapeseed is one species in which significant homoeologous exchanges have been discovered and associated with phenotypic variation [5,29,67,68]. In rapeseed, substantial homoeologous exchanges have been found and connected to phenotypic diversity. In rapeseed, where A genome replacement is more common than C genome replacement, it is proposed that directionally influenced HEs might result in sub-genome dominance [69-72], as shown in wheat [73], polyploid strawberry [74], non-crop monkey-flower [75-78], cotton [79,80] and coffee [67]. A pan-genome study of the rapeseed genome indicated two categories of gene PAV incidents: non-homoeologous exchange PAV (in which single genes differ) and homoeologous exchange PAV (in which lengthy extends of successive genes are missing caused by large genomic region exchanges) [33].

In plants, transposable elements (TEs) have been linked to genic variation development, according to pan-genome studies. The relationships between transposable elements, gene mobility, and gene PAV have been well established for many years [65,81–83]. In recent years, however, pan-genome research has provided a more refined view on the role of transposable elements on gene variability, indicating that intraspecies TEs dynamics may considerably contribute to the variation in gene addition and deletion [34,65]. It has been documented in *B. oleracea* [33] and *Brachypodium* [36,60] that TEs and varied genes are associated, and it has also been reported in rapeseed disease resistance genes [33] that TEs are associated with mutant genes [32,58]. As previously stated, a rapeseed pan-genome based on complete genomic sequences of seven populations indicated a role for different TEs in the regulation of agronomic parameters [30], which is consistent with prior results. From Barbara McClintock's discovery of genetic variants that may be changed in maize [84], TE activity has been shown to be correlated with genome rearrangement. Multiple instances of TEs that led to the gene PAV are reported, most notably in *A.*

thaliana [83] and *Z. mays* [65]. To further understand the TE–PAV relationship, additional pan-genome investigations, such as examining whether particular transposable element families are probably more linked to PAVs than others and if such links are species-specific or universal, would be helpful to researchers. It is constantly being worked on to improve the techniques for predicting and categorizing TEs within genome assemblies [85–87], which will eventually result in improved knowledge of the function of transposable elements in gene diversity in the near future.

De novo gene origin [87] is a source of mutant genes that have been relatively understudied in comparison to other sources. When the genomes of thirteen *Oryza* related species were examined, researchers discovered 185 de novo ORFs in the primary species, *Oryza japonica*, indicating that de novo gene origin is important for the creation of proteome variability in this species. Long non-coding RNAs have also been considered as a possible source of new proteins for protein synthesis [88,89] since they seem to be genetically efficient from protein-coding genes and also have higher tissue selectivity [90]. Researchers observed that non-coding transcripts were responsible for the majority of the de novo genes identified in *Oryza* [63]. In turn, comprehensive annotation and research of long non-coding RNAs (lncRNAs) may, as a result, broaden the repository of plant-specific genes and increase the likelihood of the identification of new proteins throughout the evolution process.

4. CONCLUSION AND FUTURE DIRECTIONS

The growing abundance of genome sequence data has aided pan-genome investigations, and this trend will continue as long-read sequence data quality and affordability increase rapidly. Over time, as we get a better grasp on the influence of mutant genes, it is possible that single reference assemblies may be rendered obsolete in favor of pan-genome reference assemblies. This would provide a wealth of information on genome evolution, selection, and functional properties.

Many researchers are struggling with the storage and display of pan-genome datasets. Because of the abundance of long-read sequence data, it is possible to use vg [91] or MGR [92] to store variations for whole populations. This allows the use of pan-genome mutation graphs, which record mutations for whole populations. Standards for genome structure and classification are needed to facilitate structural variance in genomes. When it comes to plant breeding populations, the use of feasible haplotype graphs for the expandable generation of pan-genomes represents a significant advancement [93,94]. In order to abstain from difficulties in arranging extremely variable and recurring sections, these graphs depend on a reference genome correlated pathway that uses genes as anchors.

A major problem is that gene and genome functional annotation methods are substantially behind methodologies for genomic assembly, and the purpose of numerous variable genes remains unsolved. However, we know that mutant genes have some characteristics in common, including that they are least likely to be syntenic, evolve under less evolutionary limitations, and have low expression levels [95,96]. According to researchers, developing a better knowledge of the roles and connections among the core and mutant genes would considerably contribute to the utility of pan-genome investigations. The use of integrative genomics approaches, which attempt to relate features of genes such as sequence integrity to their function, relationships in biological networks, and expression level [97], could be a possible approach.

Most pan-genome research conducted up till now has concentrated on the gene-containing portions of genomes; nevertheless, genomic areas outside of genes have been shown to account for a significant

fraction of phenotypic variability in plants [98]. This shows that many essential breeding parameters, such as gene PAV, may be influenced by variations in gene regulation instead of gene expression. For instance, a promoter related to fruit flavor [38,97,98] was discovered under selection in the *Solanum lycopersicum* pan-genome. With the use of epigenomic functional annotations, pan-genomes give a wealth of information on regulators that may be mined for breeding purposes.

Pan-genomes of prokaryotic organisms have recently been discovered and even crossing phylum borders, with one research including 8,203 genomes from 10 different prokaryotic phyla [95]. These investigations are computationally viable since the genomes of haploid prokaryotes are very tiny. Although no pan-genome has yet been discovered in plants, this is most likely due to computational and funding restrictions on researchers. The ability to link pan-genomes at the genus or even family level will likely become more accessible as sequencing expenditures continue to decline and computational power increases. This will allow us to think critically, like what gene content is needed to produce a legume, which will be possible as sequencing expenditures keep falling and computational power continues to rise. At last, this will enable us to anticipate and define the gene composition of every plant species, information that will have a profound impact on future genome research in general. Such comprehensive pan-genomes will enable us to address a question of centuries: which genes are responsible for the formation of a plant?

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. Proc. Natl Acad. Sci. 2005;102(39):13950-13955.
2. Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. Pangenomics comes of age: from bacteria to plant and animal applications. Trends Genet. 2020;36(2):132-145.
3. Morgante M, De Paoli E, Radovic S. Transposable elements and the plant pan-genomes. Plant Biol. 2007;10(2):149-155.
4. Golicz AA, Batley J, Edwards D. Towards plant pangenomics. Plant Biotechnol. J. 2016;14(4):1099-1105.
5. Hurgobin B, Edwards D. SNP discovery using a pangenome: has the single reference approach become obsolete? Biology. 2017;6(1):1-21.
6. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. Genom. Biol. 2020;21(1):1-9.
7. Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, Zhang SS. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. 2014;32(10):1045-1052.
8. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, Kahles A. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature. 2011;477(7365):419-423.

9. Schatz MC, Maron LG, Stein JC, Wences AH, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E, Wright MH. Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. Genome Biol. 2014;15(11):1-6.
10. Lin K, Zhang N, Severing EI, Nijveen H, Cheng F, Visser RG, Wang X, de Ridder D, Bonnema G. Beyond genomic variation-comparison and functional annotation of three *Brassica rapa* genomes: a turnip, a rapid cycling and a Chinese cabbage. BMC Genomics. 2014;15(1):1-7.
11. Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, Zhang SS. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. 2014;32(10):1045-1052.
12. Schatz MC, Maron LG, Stein JC, Wences AH, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E, Wright MH. Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. Genome Biol. 2014;15(11):1-6.
13. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, de Leon N. Insights into the maize pan-genome and pan-transcriptome. Plant Cell. 2014;26(1):121-135.
14. Yao W, Li G, Zhao H, Wang G, Lian X, Xie W. Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol. 2015;16(1):1-20.
15. Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CK, Severn-Ellis A, McCombie WR, Parkin IA, Paterson AH. The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat. Commun. 2016;7(1):1-8.
16. Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC, Zaina G, Bastien C, Cattonaro F, Marroni F, Morgante M. Characterization of the poplar pan-genome by genome-wide identification of structural variation. Mol. Biol. Evol. 2016;33(10):2706-2719.
17. Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, Stritt C, Roulin AC, Schackwitz W, Tyler L, Martin J. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nat. Commun. 2017;8(1):1-3.
18. Zhou P, Silverstein KA, Ramaraj T, Guhlin J, Denny R, Liu J, Farmer AD, Steele KP, Stupar RM, Miller JR, Tiffin P. Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. BMC Genomics. 2017;18(1):1-4.
19. Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan CK, Visendi P, Lai K, Doležel J, Batley J, Edwards D. The pangenome of hexaploid bread wheat. Plant J. 2017;90(5):1007-1013.
20. Hurgobin B, Golicz AA, Bayer PE, Chan CK, Tirnaz S, Dolatabadian A, Schiessl SV, Samans B, Montenegro JD, Parkin IA, Pires JC. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. Plant Biotechnol. J. 2018;16(7):1265-1274.
21. Ou L, Li D, Lv J, Chen W, Zhang Z, Li X, Yang B, Zhou S, Yang S, Li W, Gao H. Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence–absence variation analyses. New Phytol. 2018;220(2):360-363.
22. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T, Wang Y. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet. 2018;50(2):278-284.
23. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557(7703):43-49.

24. Yu J, Golicz AA, Lu K, Dossa K, Zhang Y, Chen J, Wang L, You J, Fan D, Edwards D, Zhang X. Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. Plant Biotechnol. J. 2019;17(5):881-892.
25. Hübner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J, Ziegler E, Lee JS, Baute GJ, Owens GL, Grassa CJ, Ebert DP. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat. Plants. 2019;5(1):54-62.
26. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat. Genet. 2019;51(6):1044-1051.
27. Song JM, Guan Z, Hu J, Guo C, Yang Z, Wang S, Liu D, Wang B, Lu S, Zhou R, Xie WZ. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. Nat. Plants. 2020;6(1):34-45.
28. Trouern-Trend AJ, Falk T, Zaman S, Caballero M, Neale DB, Langley CH, Dandekar AM, Stevens KA, Wegrzyn JL. Comparative genomics of six Juglans species reveals disease-associated gene family contractions. Plant J. 2020;102(2):410-423.
29. Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, Zhang H, Liu Z, Shi M, Huang X. Pan-genome of wild and cultivated soybeans. Cell. 2020;182(1):162-176.
30. Song JM, Guan Z, Hu J, Guo C, Yang Z, Wang S, Liu D, Wang B, Lu S, Zhou R, Xie WZ. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. Nat. Plants. 2020;6(1):34-45.
31. Jiao WB, Schneeberger K. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat. Commun. 2020;11(1):1-10.
32. Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CK, Severn-Ellis A, McCombie WR, Parkin IA, Paterson AH. The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat. Commun. 2016;7(1):1-8.
33. Hurgobin B, Golicz AA, Bayer PE, Chan CK, Tirnaz S, Dolatabadian A, Schiessl SV, Samans B, Montenegro JD, Parkin IA, Pires JC. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. Plant. Biotechnol. J. 2018;16(7):1265-1274.
34. Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, Stritt C, Roulin AC, Schackwitz W, Tyler L, Martin J. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nat. Commun. 2017;8(1):1-3.
35. Yu J, Golicz AA, Lu K, Dossa K, Zhang Y, Chen J, Wang L, You J, Fan D, Edwards D, Zhang X. Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. Plant Biotechnol. J. 2019;17(5):881-892.
36. Zhao J, Bayer PE, Ruperao P, Saxena RK, Khan AW, Golicz AA, Nguyen HT, Batley J, Edwards D, Varshney RK. Trait associations in the pangenome of pigeon pea (*Cajanus cajan*). Plant Biotechnol. J. 2020;18(9):1946-1954.
37. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T, Wang Y. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet. 2018;50(2):278-284.
38. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat. Genet. 2019;51(6):1044-1051.
39. Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, Zhang H, Liu Z, Shi M, Huang X. Pan-genome of wild and cultivated soybeans. Cell. 2020;182(1):162-176.
40. Sears ER, Miller TE. The history of Chinese Spring wheat. Cereal Res. Commun. 1985;261-263.

41. Nsabiyera V, Baranwal D, Qureshi N, Kay P, Forrest K, Valárik M, Doležel J, Hayden MJ, Bariana HS, Bansal UK. Fine mapping of Lr49 using 90K SNP chip array and flow-sorted chromosome sequencing in wheat. Front. Plant Sci. 2020;1787.
42. Tian X, Li R, Fu W, Li Y, Wang X, Li M, Du D, Tang Q, Cai Y, Long Y, Zhao Y. Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. Sci. China Life Sci. 2020;63(5):750-763.
43. Li R, Fu W, Su R, Tian X, Du D, Zhao Y, Zheng Z, Chen Q, Gao S, Cai Y, Wang X. Towards the complete goat pan-genome by recovering missing genomic segments from the reference genome. Front. Genet. 2019;10:1169.
44. Pimentel D, Wilson C, McCullum C, Huang R, Dwen P, Flack J, Tran Q, Saltman T, Cliff B. Economic and environmental benefits of biodiversity. BioScience. 1997;47(11):747-757.
45. Doebley JF, Gaut BS, Smith BD. The molecular genetics of crop domestication. Cell. 2006;127(7):1309-1321.
46. Schouten HJ, Tikunov Y, Verkerke W, Finkers R, Bovy A, Bai Y, Visser RG. Breeding has increased the diversity of cultivated tomato in The Netherlands. Front. Plant Sci. 2019;1606.
47. Tian D, Traw MB, Chen JQ, Kreitman M, Bergelson J. Fitness costs of R-gene-mediated resistance in *Arabidopsis thaliana*. Nature. 2003;423(6935):74-77.
48. Kehr B, Helgadottir A, Melsted P, Jonsson H, Helgason H, Jonasdottir A, Jonasdottir A, Sigurdsson A, Gylfason A, Halldorsson GH, Kristmundsdottir S. Diversity in non-repetitive human sequences not found in the reference genome. Nat. Genet. 2017;49(4):588-593.
49. Manni M, Zdobnov E. Microbial contaminants cataloged as novel human sequences in recent human pan-genomes. Nature. 2012;127(7):133-139.
50. Van de Weyer AL, Monteiro F, Furzer OJ, Nishimura MT, Cevik V, Witek K, Jones JD, Dangl JL, Weigel D, Bemm F. A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. Cell. 2019;178(5):1260-1272.
51. Pryor T. The origin and structure of fungal disease resistance genes in plants. Trends Genet. 1987;3:157-161.
52. Crute IR, Pink D. Genetics and utilization of pathogen resistance in plants. The Plant Cell. 1996;8(10):1747.
53. Michelmore RW, Meyers BC. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. 1998;8(11):1113-1130.
54. Shi J, Zhang M, Zhai W, Meng J, Gao H, Zhang W, Han R, Qi F. Genome-wide analysis of nucleotide binding site-leucine-rich repeats (NBS-LRR) disease resistance genes in *Gossypium hirsutum*. Physiol. Mol. Plant P. 2018;104:1-8.
55. Leister D, Kurth J, Laurie DA, Yano M, Sasaki T, Devos K, Graner A, Schulze-Lefert P. Rapid reorganization of resistance gene homologues in cereal genomes. Proc. Natl Acad. Sci. 1998;95(1):370-375.
56. Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, Diers BW. Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science. 2012;338(6111):1206-1269.
57. Chae E, Bomblies K, Kim ST, Karelina D, Zaidem M, Ossowski S, Martín-Pizarro C, Laitinen RA, Rowan BA, Tenenboim H, Lechner S. Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. Cell. 2014;159(6):1341-1351.
58. Bayer PE, Golicz AA, Tirnaz S, Chan CK, Edwards D, Batley J. Variation in abundance of predicted resistance genes in the *Brassica oleracea* pangenome. Plant Biotechnol. J. 2019;17(4):789-800.

59. Dolatabadian A, Bayer PE, Tirnaz S, Hurgobin B, Edwards D, Batley J. Characterization of disease resistance genes in the *Brassica napus* pangenome reveals significant structural variation. Plant Biotechnol. J. 2020;18(4):969-982.
60. Sudupak MA, Bennetzen JL, Hulbert SH. Unequal exchange and meiotic instability of disease-resistance genes in the Rp1 region of maize. Genetics. 1993;133(1):119-125.
61. Kuang H, Woo SS, Meyers BC, Nevo E, Michelmore RW. Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. The Plant Cell. 2004;16(11):2870-2894.
62. Panchy N, Lehti-Shiu M, Shiu SH. Evolution of gene duplication in plants. Plant Physiol. 2016;171(4):2294-2316.
63. Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, Wen B. Rapid evolution of protein diversity by de novo origination in *Oryza*. Nat. Ecol. Evol. 2019;3(4):679-690.
64. Dunning LT, Olofsson JK, Parisod C, Choudhury RR, Moreno-Villena JJ, Yang Y, Dionora J, Quick WP, Park M, Bennetzen JL, Besnard G. Lateral transfers of large DNA fragments spread functional genes among grasses. Proc. Natl Acad. Sci. 2019;116(10):4416-4425.
65. Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. PLoS Biol. 2010;8(6):10-40.
66. Woodhouse MR, Pedersen B, Freeling M. Transposed genes in Arabidopsis are often associated with flanking repeats. PLoS Genet. 2010;6(5):42-50.
67. Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, Smith RD, Teresi SJ, Nelson AD, Wai CM, Alger EI. Origin and evolution of the octoploid strawberry genome. Nat. Genet. 2019;51(3):541-547.
68. Bird KA, Niederhuth CE, Ou S, Gehan M, Pires JC, Xiong Z, VanBuren R, Edger PP. Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid *Brassica napus*. New Phytol. 2021;230(1):354-371.
69. Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC. Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. Genetics. 2012;190(4):1563-1574.
70. Cheng F, Wu J, Wang X. Genome triplication drove the diversification of *Brassica* plants. Hortic. Res. 2014;1.
71. Golicz AA. Construction and analysis of the *Brassica oleracea* pangenome. Plant Biotechnol. J. 2019;17(4):689-700.
72. Bird KA, VanBuren R, Puzey JR, Edger PP. The causes and consequences of subgenome dominance in hybrids and recent polyploids. New Phytol. 2018;220(1):87-93.
73. Chalhoub B, Denoeud F, Liu S, Parkin IA, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, Corréa M. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science. 2014;345(6199):950-953.
74. Samans B, Chalhoub B, Snowdon RJ. Surviving a genome collision: genomic signatures of allopolyploidization in the recent crop species *Brassica napus*. Plant Genome. 2017;10(3):1-5.
75. Feldman M, Levy AA, Fahima T, Korol A. Genomic asymmetry in allopolyploid plants: wheat as a model. J. Exp. Bot. 2012;63(14):5045-5059.
76. Appels R, Eversole K, Stein N, Feuillet C, Keller B, Rogers J, Pozniak CJ, Choulet F, Distelfeld A, Poland J. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 2018;361(6403):71-91.
77. Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, Davey M, Jacobs J, Van Ex F, Pasha A, Khedikar Y. The transcriptional landscape of polyploid wheat. Science. 2018;361(6403):60-72.

78. Bardil A, de Almeida JD, Combes MC, Lashermes P, Bertrand B. Genomic expression dominance in the natural allopolyploid *Coffea arabica* is massively affected by growth temperature. New Phytol. 2011;192(3):760-774.
79. Yoo MJ, Szadkowski E, Wendel JF. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity. 2013;110(2):171-180.
80. Edger PP, Smith R, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y, Bewick AJ, Ji L, Platts AE, Bowman MJ, Childs KL. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. Plant Cell. 2017;29(9):2150-2167.
81. Kashkush K, Feldman M, Levy AA. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics. 2002;160(4):1651-1659.
82. Hawkins JS, Proulx SR, Rapp RA, Wendel JF. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proc. Natl Acad. Sci. 2009;106(42):17811-17816.
83. Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. Genome Res. 2008;18(12):1924-1937.
84. McClintock B. Induction of instability at selected loci in maize. Genetics. 1953;38(6):579.
85. Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, Lugo CS, Elliott TA, Ware D, Peterson T, Jiang N. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20(1):1-8.
86. Yan H, Bombarely A, Li S. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics. 2020;36(15):4269-4275.
87. Cruz MH, Domingues DS, Saito PT, Paschoal AR, Bugatti PH. TERL: classification of transposable elements by convolutional neural networks. Briefings in bioinformatics. 2021;22(3):185-198.
88. Van Oss SB, Carvunis AR. De novo gene birth. PLoS Gen. 2019;15(5):100-130.
89. Golicz AA, Bhalla PL, Singh MB. lncRNAs in plant and animal sexual reproduction. Trends Plant Sci. 2018;23(3):195-205.
90. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. Life. 2014;33(3):35-43.
91. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. 2018;36(9):875-879.
92. Rabbani L, Müller J, Weigel D. An algorithm to build a multi-genome reference. bioRxiv. 2020;23(4):25-33.
93. Jensen SE, Charles JR, Muleta K, Bradbury PJ, Casstevens T, Deshpande SP, Gore MA, Gupta R, Ilut DC, Johnson L, Lozano R. A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. Plant Genome. 2020;13(1):20-39.
94. Contreras-Moreira B, Cantalapiedra CP, García-Pereira MJ, Gordon SP, Vogel JP, Igartua E, Casas AM, Vinuesa P. Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. Front. Plant Sci. 2017;8:184.
95. Golicz AA, Bhalla PL, Singh MB. MCRiceRepGP: a framework for the identification of genes associated with sexual reproduction in rice. Plant J. 2018;96(1):188-202.
96. Hassani-Pak K, Castellote M, Esch M, Hindle M, Lysenko A, Taubert J, Rawlings C. Developing integrated crop knowledge networks to advance candidate gene discovery. Appl. Transl. Genom. 2016;11:18-26.

97. Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. Open chromatin reveals the functional maize genome. Proc. Natl Acad. Sci. 2016;113(22):3177-3184.
98. Maistrenko OM, Mende DR, Luetge M, Hildebrand F, Schmidt TS, Li SS, Rodrigues JF, von Mering C, Pedro Coelho L, Huerta-Cepas J, Sunagawa S. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. ISME J. 2020;14(5):1247-1259.