Title: Plant pan-genomes are the new reference

Authors: Philipp E. Bayer[1], Agnieszka A. Golicz[2], Armin Scheben[3], Jacqueline Batley[1], David Edwards[1*]

Affiliations:

1. School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, WA, Australia
2. Plant Molecular Biology and Biotechnology Laboratory, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, Melbourne, VIC, Australia
3. Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, United States of America

Contact information: dave.edwards@uwa.edu.au

Abstract

Recent years have seen a surge in plant genome sequencing projects and the comparison of multiple related individuals. The high degree of genomic variation observed led to the realisation that single reference genomes do not represent the diversity within a species, and led to the expansion of the pan-genome concept. Pan-genomes represent the genomic diversity of a species and includes core genes, found in all individuals, as well as variable genes which are absent in some individuals. Variable gene annotations often show similarities across plant species, with genes for biotic and abiotic stress commonly enriched within variable gene groups. Here we review the growth of pan-genomics in plants, explore the origins of gene presence/absence variation and show how pan-genomes can support plant breeding and evolution studies.

Pan-genomes in plants: beginnings and current status

The concept of pan-genomes was first developed in bacteria in 2005[1], where the sequencing of several isolates of *Streptococcus agalactiae* revealed a core genome represented by 80% of *S. agalactiae* genes, with the other 20% being absent in at least one isolate[1]. However, it took almost 10 years for plant pan-genomes to be constructed after the initial bacterial pan-genome work. This was partially due to the expense of data generation, but also the expectation that there would be very little gene presence/absence variation (PAV) in higher organisms which do not exchange genetic material as freely as bacteria[2]. The first publication to apply the term pan-genome in plants appeared in 2007, where it described short variable regions in the rice and maize genomes[3]. However, the extent of gene presence/absence was not understood at that time due to lack of accurate whole genome assemblies for multiple individuals of the same species. However, as DNA sequencing costs declined, it became feasible to undertake whole genome comparisons within species, and three general approaches for pan-genome assembly were developed[4,5] (Figure 1). The first method developed was the whole genome assembly and comparison, where the genomes of multiple individuals are assembled and then compared. This was later complemented by the iterative assembly and presence/absence variation calling approach, where genomic reads from multiple individuals are aligned to a reference, and non-aligning reads assembled and added to the growing pan-genome reference. Subsequent remapping of all reads to the pan-genome permits PAV calling across the population. More recently there have been rapid developments in graph based pan-genome assembly, where a graph representing genomic diversity and conservation is constructed[6].

The first two approaches are highly complementary, with whole genome assembly and comparison providing important structural and gene position information, while the iterative assembly approach permits extension to very large numbers of individuals, the identification of rare genes, and the distribution of presence/absence variation in a population. The graph assembly approach has been used routinely in bacteria but is only now becoming feasible in more complex genomes due to improvements in long read DNA sequencing.

Each of these approaches carries their own benefits and drawbacks: for example, iterative assembly does not differentiate between extreme sequence divergence at a locus and the insertion/deletion of sequences, while the whole genome assembly approach cannot differentiate between real genome diversity between individuals and the common errors and variation observed in assembly and annotation methods. Graph based pan-genomes are likely to represent the future of plant pan-genomics, however the requirement of significant quantities of high quality long read data, as well as methodological constraints in the form of large computational memory requirements limits their current application to relatively few individuals and small genomes.

There has been a rapid growth of plant pan-genome studies (Figure 2, Table 1). The first published plant pan-genome was based on a comparison of whole genome assemblies for seven wild soybean individuals[7] and found variable genes linked with seed composition, flowering and maturity time, organ size and biomass, and additional copies of disease resistance genes in the wild *Glycine soja* which are not present in domesticated *G. max.* An earlier study examined 18 *Arabidopsis thaliana* accessions, but mostly compared gene expression and protein isoforms rather than gene presence/absence[8]. At the same time as the soybean whole genome assemblies, another publication examined a small rice pan-genome based on three divergent accessions[9], finding a deletion in the *S5* hybrid sterility locus in one cultivar, and presence/absence variation in the submergence tolerance gene *Sub1A*.

Subsequent whole genome assembly and comparison studies include the recent comparison of eight *de novo* assembled *B. napus* genomes[10], where two PAVs were associated with flowering time, with both representing hAT retrotransposon insertions within known flowering time genes. Another study in *A. thaliana* based on seven assemblies found non-syntenic hotspots of rearrangements (HOTs) associated with tandem duplications[11]. These HOTs undergo reduced meiotic recombination, contain fewer genes, and are significantly enriched for disease resistance genes. Due to the absence of high quality genome assemblies for several individuals of a species, mining for HOTs is currently not possible in other plant species, and any link between HOTs and variable disease resistance remains to be validated.

Due to the cost of generating high quality assemblies for multiple individuals, several pan-genome studies have applied the iterative mapping and assembly approach[4], including *Brassica oleracea*, based on 10 individuals[12], bread wheat based on 18 individuals[13], and canola based on 53 individuals[14]. These early plant pan-genome studies produced two major findings, that there is a large variable gene content in each of the species studied (15 to 40%), and that the genes which display presence/absence variation are frequently annotated with predicted functions associated with biotic and abiotic stress tolerance.

Subsequent studies include a pan-genome representing 54 *Brachypodium distachyon* individuals which identified an additional 7,135 genes, and found that some variable genes are core within subpopulations, thereby perpetuating population structure[15]. A pan-genome of sesame (*Sesamum indicum*) was assembled using five individuals[16] which enabled a genomic comparison between old and modern sesame cultivars, suggesting that pan-genomes can help track genes which change in frequency during domestication and breeding. In a recent pan-genome study of 89 pigeon pea (*Cajanus cajan*) individuals[17], the variable genes were used for association analysis and three genes were found to be associated with seed weight, suggesting that PAVs can complement SNPs for trait association.

With the growing interest in pan-genome studies, there has been an increased interest in the distribution of variable genes in populations. For example, a rice pan-genome study examined gene variation across 66 representatives from a collection of 1,083 *Oryza sativa* and 446 wild *O. rufipogon* accessions[18] and found 10,783 newly identified genes that were at least partially missing in the reference assembly. These included genes previously shown to be agronomically important; linked with submergence tolerance and phosphorus deficiency tolerance, confirming earlier observations based on three rice accessions[9]. A pan-genome study in tomato examined variation across 725 diverse lines[19], and identified an additional 4,873 genes, many of which were linked with disease resistance, along with a rare allele linked with tomato flavour which was selected against during domestication, but reappeared in modern tomato cultivars due to wild introgressions. A recent pan-genome study in soybean combined assembly of 26 lines with resequencing data from a diverse set of 2,896 to examine diversity associated with agronomic traits[20].

Impact of pan-genomes on plant biology

The individuals selected for reference genome sequencing are frequently chosen for historic reasons, for example in bread wheat, Chinese Spring was selected because the standard karyotype system was developed using this cultivar[21]. However the gene content of Chinese Spring was found to be very different from modern varieties, with the first bread wheat pan-

genome study finding 12,150 genes present in all 18 re-sequenced modern varieties but absent from Chinese Spring[13]. Use of a single reference can impact our understanding of the genomic basis of traits, for example, the wheat rust resistance gene *Lr49* shows significant structural variation between varieties[22]. Moving to the use of pan-genomes as references will improve a broad range of genomic analysis. For example, using a pan-genome reference improves short read mapping accuracy compared to using a single reference, resulting in higher quality variant calls and more accurate gene expression quantification[12,23,24]. Delineating plant species based on gene content remains a challenge, particularly considering the significant gene presence/absence variation between individuals of a species. However, as pan-genomes are developed for an increasing number of species, our greater understanding of gene conservation and loss may assist in defining species level differences in gene content.

Beyond basic biology, an understanding of gene presence/absence variation can support applications for crop improvement. Crop wild relatives (CWRs) often contain a broader repertoire of genes and provide a valuable source of genetic diversity which can be applied for crop breeding, with as much as 30% of the increases in crop yield during the late 20th century being attributed to the use of CWRs in plant breeding programs[25]. Pangenomic analysis allows us to examine gene retention and loss during domestication and breeding[26], supporting the characterisation of lost diversity and the potential to reintroduce genes into modern varieties. For example, gene loss linked with flavour which occurred during tomato domestication in South America and Mesoamerica has since been reintroduced into modern cultivars[19,27]. Studies of gene distribution across wild species in different environments could support the breeding of crop plants with greater adaptation to diverse environments and resilience to climate change.

Functional analysis of variable genes across plant pan-genomes suggests that they are enriched for those related to responses to abiotic and biotic stress, especially disease resistance, as there is a fitness cost associated with disease resistance genes[28]. The variability of disease resistance genes is observed in monocots, such as wheat[13], as well as in dicots such as *B. napus*[14], *B. oleracea*[12] and tomato[19], and similar observations are even being reported in the human pan-genome[29,30]. These observations have led to the concept of the pan-NLRome, a pan-genome study that focuses exclusively on nucleotide binding leucine rich repeat receptor (NLR) disease resistance genes[31], which so far has only been carried out in *A. thaliana* where it was found that just 37 out of 64 accessions were sufficient to recover 90% of the predicted NLR gene content. Disease resistance genes are often organised in tightly linked physical clusters[32-35], some of which are highly variable[36-40]. Differences between

clustered and unclustered disease resistance genes are likely due to unequal crossing over and meiotic instability caused by paralogous copies in these clusters[41] which occurs only for some groups of disease resistance genes (type I), while type II disease resistance genes show infrequent genomic changes in *B. oleracea*[42], similar to observations in *A. thaliana*[31].

While the variable gene fraction tends to be enriched for disease resistance genes, this is not the case in all plant species. For example, the *Amborella* pan-genome contains relatively few disease resistance genes and these tend to be core genes (Haifei Hu, personal communication), a fact which may reflect the unusual geographic location and evolutionary history of this species. Variable genes are also often associated with abiotic stress and environmental adaptation[12,13], suggesting that these genes may support future crop breeding strategies.

Origin of variable genes in plants

While the prevalence of gene presence/absence variation is firmly established, the origin of variable genes is relatively poorly understood. Several mechanisms of gene gain and loss, which could contribute to variable gene generation, have been described in plants (Figure 3). New genes can be gained via whole genome duplications (WGDs), local tandem duplications, TE mediated duplications, segmental duplications, introgression from related species, horizontal gene transfer and *de novo* gene birth[43-45]. Genes can also be lost due to deletions, for example mediated by intra-chromosomal recombination and pseudogenization[43,46,47]. Analysis of plant pan-genomes allows a more complete picture of the relative contributions of different mechanisms of gain and loss to the overall species gene content, and provide an understanding of how selection may lead to changes in the frequency of variable genes.

Polyploid species seem to have a greater proportion of variable genes than diploids, however there are currently few polyploid pan-genomes available to confirm this trend. Variable gene content in allopolyploids can be shaped by dominant sub-genomes, as observed between sub-genomes in strawberry[48] and canola[10,14,49], with dominant sub-genomes hosting a greater proportion of core genes. Whole genome duplications result in doubling of the entire gene complement and are often followed by gene loss, also known as fractionation. For example, the *Brassica* lineage underwent a whole genome triplication (WGT) event followed by differential fractionation, resulting in three sub-genomes (LF – least fractionated, MF1 – more fractionated 1 and MF2 – more fractionated 2)[50,51]. Subsequent analysis of the *B. oleracea* pan-genome revealed a significant association between the sub-genome assignment and the proportion of variable genes, with LF harbouring the least and MF2 the most variable genes[52]. The sub-genomic location of variable genes in *Brassica* likely correlates with rate of gene loss,

even on very short evolutionary scales, and impacts intraspecies variation. However, the majority of *B. oleracea* variable genes were not assigned to sub-genomes[52], reflecting a similar observation in *Brachypodium,* where variable genes were shown to be less syntenic with orthologous regions of other grasses, suggesting that they are more likely to evolve outside of syntenic blocks[15]. A pan-genome study of sesame, which underwent a WGD approximately 70 MYA, aimed to identify the origin of core and variable genes. Almost half of the core genes and only ~10% of the variable genes could be traced to WGD. The low proportion of variable genes attributed to WGD origin reflects that many are not found in syntenic blocks. A similar proportion of the core and variable genes (~10%) could be assigned to local tandem duplications, suggesting that for sesame, tandem duplications are not a significant source of variable genes, although line specific variations do exist[16].

Homoeologous exchange (HE) in amphipolyploid plants are another common cause of gene presence/absence variation[14,53]. *Brassica napus* is one species in which extensive HEs have been observed and linked to phenotypic variation[14,49,54,55]. It has been suggested that directionally biased homoeologous exchanges, such as observed in *B. napus,* where replacement of A genome with C genome is more common, can lead to sub-genome dominance[53], which has been observed for several other species including polyploid strawberry[48], wheat[56-58], coffee[59], and cotton[60], as well as the non-crop species monkeyflower[61]. Analysis of the *B. napus* pan-genome revealed two types of gene PAV events: non-HE PAV, where individual genes are variable, and HE-PAV where longer stretches of consecutive genes are absent due to the exchange of large genomic segments[14].

Plant pan-genome studies have highlighted the role of transposable elements (TEs) in the generation of genic diversity. The link between TEs, gene presence/absence variation and gene movement has been long acknowledged[46,62-64], however recent pan-genome studies have enabled a more fine grained view of the impact of TEs on gene variability, suggesting that intraspecies TE dynamics could be an important contributor to variable gene birth and loss[15,46]. Variable *B. napus* disease resistance genes identified by Hurgobin, et al. [14] tend to be co-located with transposable elements[39], and a similar association between TEs and variable genes was observed in *Brachypodium*[15] and *B. oleracea*[12,39]. Additionally, a *B. napus* pan-genome based on whole genome assembly of eight individuals has revealed a role of variable TEs in agronomic traits[10]. TE activity has been associated with genome reshuffling since Barbara McClintock's discovery of moveable genetic elements in maize[65], and since then, many examples of TEs leading to gene PAV have been discovered, including in *Arabidopsis*[64] and maize[46]. Future pan-genome studies will assist our understanding of this TE-PAV association, for example investigating whether certain TE families are more likely to

be associated with PAV and whether these relationships are universal or species specific. Methods to predict and classify TEs in genome assemblies are constantly improving[66-68] and this will provide a greater understanding of the role of TEs in gene variability.

One relatively underexplored source of variable genes is *de novo* gene birth[69]. A recent comparative analysis of 13 closely related *Oryza* genomes identified 175 *de novo* open reading frames in the focal species *O. sativa* subspecies *japonica*, providing support for the role of *de novo* gene birth in the generation of proteome diversity. It has also been suggested that long non-coding RNAs (lncRNAs), which tend to be evolutionarily younger than protein coding genes and display higher tissue specificity[70], may provide a reservoir for the synthesis of new proteins[71]. In *Oryza*, 91% of the *de novo* genes identified originated from non-coding transcripts[44]. Detailed annotation and analysis of lncRNAs may therefore extend the repertoire of plant variable genes and provide candidates for the identification of newly evolving proteins.

Prospects and future directions

Pan-genome studies have been supported by the increasing availability of genome sequence data, and this will continue with the rapid improvements in the quality and the reducing cost of long read sequence data. An increase in understanding the impact of variable genes may lead to single reference assemblies becoming redundant, with the pan-genome increasingly becoming the new reference, providing broad insights into evolution, selection and in particular the functionality of genomes.

One of the challenges of pan-genome analysis is the storage and visualisation of pan-genome data. Abundant long read sequence data supports the application of pan-genome variation graphs, which store variations for entire populations, such as implemented in vg[72] or MGR[73]. There is a need to establish standards for genome structure and annotation which accommodate structural genome variation. In plant breeding populations, a step forward is the use of practical haplotype graphs for the scalable construction of pan-genomes[74]. These graphs rely on a reference genome coordinate system and use genes to anchor sequences, allowing them to avoid challenges in aligning repetitive and highly divergent regions.

Methods for the accurate and consistent functional annotation of genes and genomes are significantly lagging behind approaches for their assembly, and the role of many variable genes remains unknown. We do know that variable genes share certain properties, including being less likely to be syntenic, evolving under reduced evolutionary constraints and having lower expression levels[12,15,75]. Improving our understanding of the functions and interactions between the core and variable genes will significantly add value to pan-genome studies. One

possible approach involves integrative genomics methodologies, which aim to link properties of genes such as expression level, connectivity in biological networks, and sequence conservation to their function[76,77] to gain a broader understanding of their potential function.

The majority of pan-genome studies to date have focussed on the genic portion of the genome, however genomic regions outside of genes explain a substantial proportion of phenotypic variance in crops[78]. This suggests that many important agronomic traits may be determined by changes in gene regulation rather than gene PAV. For example, regions found to be under selection in the tomato pan-genome include a promoter associated with fruit flavour[19]. Combined with epigenomic functional annotations of regulators, pan-genomes provide a rich resource to mine for regulatory sequence variation that can be harnessed in breeding.

Recently, prokaryotic pan-genomes have crossed species and even phyla boundaries, including one study using 7,104 genomes from ten prokaryotic phyla[79]. Due to the small size of haploid prokaryotic genomes such studies are computationally feasible. In plants however, currently no pan-genome has crossed the genus boundary, likely due to computational and financial constraints. As sequencing costs continue to fall and computational power rises, plant pan-genome studies are likely to expand beyond the species level, to where we can start to connect pan-genomes on the genus, or even the family level, allowing us to ask questions such as what gene content is required to make a legume, and eventually being able to predict and characterise the gene content of all plant species, knowledge which will revolutionise future genome studies. Such wide pan-genomes will allow us to answer an age old question, what genes make a plant?

Table 1. Summary of plant pan-genome studies

Year published	Approach	Species	Domestication status	Ploidy	Number of accessions	Pangenome genes	Core genome (% genes / gene clusters)	Number of dispensable genes / gene clusters missing from reference (cultivar)	Refs
2014	de novo	*Brassica rapa*	crop	Diploid	3	41,858 genes	87	2,830 (Chiifu)	80
	de novo	*Glycine soya (soybean)*	wild	Tetraploid (diploidized)	7	59,080 gene clusters	49	NA	81
	de novo	*Oryza sativa*	crop	Diploid	3	39,891 genes	92	1,300 (Nipponbare)	82
	de novo transcriptome	*Zea mays (maize)*	crop	Tetraploid (diploidized)	503	41,903 transcripts	39	8,681 (B73)	83
2015	de novo (metagenome assembly)	*Oryza sativa (indica/japonica)*	crop	Diploid	1483	NA	Additional genes found (8991/6366)	8,000 (Nipponbare)	84
2016	iterative assembly	*Brassica oleracea*	crop	Diploid	10	61,379 genes	81	6,922 (TO1000)	85
	read mapping (no assembly)	*Populus (poplar)*	wild	Diploid	7	NA	<90	NA	86
2017	de novo	*Brachypodium distachyon (stiff brome)*	wild	Diploid	54	37,886 genes	55	7,135 (Bd21)	87
	de novo	*Medicago truncatula*	wild	Diploid	15	75,000 gene clusters	33	38,000 (HM101)	88
	iterative assembly	*Triticum aestivum (bread wheat)*	crop	Hexaploid	19	139,747 genes	64	12,150 (Chinese Spring)	13

2018	iterative assembly	*Brassica napus*	crop	Tetraploid	53	94,013 genes	62	13,633 (Darmor)	89
	iterative assembly	*Capsicum (pepper)*	crop	Diploid	383	51,757 genes	56	6,984 (Zunla-1)	90
	iterative assembly	*Oryza sativa/Oryza rufipogon*	crop	Diploid	67	42,580 genes	62	10,872 (Nipponbare)	91
	map-to-pan	*Oryza sativa (rice)*	crop	Diploid	3010	48,098 genes	54–62	12,465 (Nipponbare)	92
2019	de novo	*Sesamum indicum (sesame)*	under-utilized crop	Diploid	5	26,472 gene clusters	58	3,084 (Zhongzhi13)	93
	iterative assembly	*Helianthus annuus (sunflower)*	crop	Diploid	493	61,205 genes	83	17,061 (HA412-HO)	94
	iterative assembly	*Solanum lycopersicum (tomato)*	crop	Diploid	725	40,369 genes	74	4,873 (Heinz 1706)	95
2020	de novo	*Brassica napus (oilseed rape)*	crop	Tetraploid	9	105,672 gene clusters	56	5,912 (Darmor)	96
	de novo	*Juglans (walnut)*	wild	Diploid	6	26,458 gene clusters	55	NA	97
	de novo, graph	*Glycine max (soybean)*	crop	Diploid	29	57,492 gene clusters	50	26,676 (ZH13)	98

References

1 Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**, 13950-13955, doi:10.1073/pnas.0506758102 (2005).

2 Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J. & Edwards, D. Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends in Genetics* **36**, 132-145, doi:https://doi.org/10.1016/j.tig.2019.11.006 (2020).

3 Morgante, M., De Paoli, E. & Radovic, S. Transposable elements and the plant pan-genomes. *Curr Opin Plant Biol* **10**, 149-155, doi:10.1016/j.pbi.2007.02.001 (2007).

4 Golicz, A. A., Batley, J. & Edwards, D. Towards plant pangenomics. *Plant Biotechnol J* **14**, 1099-1105, doi:10.1111/pbi.12499 (2016).

5 Hurgobin, B. & Edwards, D. SNP discovery using a pangenome: has the single reference approach become obsolete? *Biology* **6**, 21 (2017).

6 Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs. *arXiv* (2020).

7 Li, Y. H. *et al.* De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* **32**, 1045-1052, doi:10.1038/nbt.2979 (2014).

8 Gan, X. *et al.* Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**, 419-423, doi:10.1038/nature10414 (2011).

9 Schatz, M. C. *et al.* Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol* **15**, 506, doi:10.1186/PREACCEPT-2784872521277375 (2014).

10 Song, J.-M. *et al.* Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants*, 1-12 (2020).

11 Jiao, W.-B. & Schneeberger, K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *BioRxiv*, 738880 (2019).

12 Golicz, A. A. *et al.* The pangenome of an agronomically important crop plant Brassica oleracea. *Nat Commun* **7**, 13390, doi:10.1038/ncomms13390 (2016).

13 Montenegro, J. D. *et al.* The pangenome of hexaploid bread wheat. *The Plant Journal* **90**, 1007-1013 (2017).

14 Hurgobin, B. *et al.* Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. *Plant Biotechnol J* **16**, 1265-1274, doi:10.1111/pbi.12867 (2018).

15 Gordon, S. P. *et al.* Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nature communications* **8**, 2184 (2017).

16 Yu, J. *et al.* Insight into the evolution and functional characteristics of the pan‐genome assembly from sesame landraces and modern cultivars. *Plant biotechnology journal* **17**, 881-892 (2019).

17 Zhao, J. *et al.* Trait associations in the pangenome of pigeon pea (Cajanus cajan). *Plant biotechnology journal* (2020).

18 Zhao, Q. *et al.* Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet* **50**, 278-284, doi:10.1038/s41588-018-0041-z (2018).

19 Gao, L. *et al.* The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet* **51**, 1044-1051, doi:10.1038/s41588-019-0410-2 (2019).

20 Liu, Y. *et al.* Pan-Genome of Wild and Cultivated Soybeans. *Cell*, doi:https://doi.org/10.1016/j.cell.2020.05.023 (2020).

21 Sears, E. & Miller, T. The history of Chinese Spring wheat. *Cereal Research Communication*, 261-263 (1985).

22 Nsabiyera, V. *et al.* Fine mapping of Lr49 using 90K SNP chip array and flow sorted chromosome sequencing in wheat. *Frontiers in Plant Science* **10**, 1787 (2019).
23 Tian, X. *et al.* Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. *Science China Life Sciences*, doi:10.1007/s11427-019-9551-7 (2019).
24 Li, R. *et al.* Towards the Complete Goat Pan-Genome by Recovering Missing Genomic Segments From the Reference Genome. *Frontiers in Genetics* **10**, doi:10.3389/fgene.2019.01169 (2019).
25 Pimentel, D. *et al.* Economic and Environmental Benefits of Biodiversity. *BioScience* **47**, 747-757, doi:10.2307/1313097 (1997).
26 Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127**, 1309-1321, doi:10.1016/j.cell.2006.12.006 (2006).
27 Schouten, H. J. *et al.* Breeding has increased the diversity of cultivated tomato in The Netherlands. *Frontiers in Plant Science* **10**, 1606 (2019).
28 Tian, D., Traw, M., Chen, J., Kreitman, M. & Bergelson, J. Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. *Nature* **423**, 74-77 (2003).
29 Kehr, B. *et al.* Diversity in non-repetitive human sequences not found in the reference genome. *Nature Genetics* **49**, 588 (2017).
30 Manni, M. & Zdobnov, E. M. Microbial contaminants cataloged as novel human sequences in recent human pan-genomes. *bioRxiv* (2020).
31 Van de Weyer, A.-L. *et al.* A species-wide inventory of NLR genes and alleles in Arabidopsis thaliana. *Cell* **178**, 1260-1272. e1214 (2019).
32 Pryor, T. The origin and structure of fungal disease resistance genes in plants. *Trends in Genetics* **3**, 157-161 (1987).
33 Crute, I. R. & Pink, D. Genetics and utilization of pathogen resistance in plants. *The Plant Cell* **8**, 1747 (1996).
34 Michelmore, R. W. & Meyers, B. C. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res* **8**, 1113-1130, doi:10.1101/gr.8.11.1113 (1998).
35 Shi, J. *et al.* Genome-wide analysis of nucleotide binding site-leucine-rich repeats (NBS-LRR) disease resistance genes in Gossypium hirsutum. *Physiological and Molecular Plant Pathology* **104**, 1-8 (2018).
36 Leister, D. *et al.* Rapid reorganization of resistance gene homologues in cereal genomes. *Proceedings of the National Academy of Sciences* **95**, 370-375 (1998).
37 Cook, D. E. *et al.* Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *science* **338**, 1206-1209 (2012).
38 Chae, E. *et al.* Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. *Cell* **159**, 1341-1351 (2014).
39 Bayer, P. E. *et al.* Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome. *Plant Biotechnol J* **17**, 789-800, doi:10.1111/pbi.13015 (2019).
40 Dolatabadian, A. *et al.* Characterization of disease resistance genes in the Brassica napus pangenome reveals significant structural variation. *Plant biotechnology journal* (2019).
41 Sudupak, M. A., Bennetzen, J. & Hulbert, S. H. Unequal exchange and meiotic instability of disease-resistance genes in the Rp1 region of maize. *Genetics* **133**, 119-125 (1993).
42 Kuang, H., Woo, S.-S., Meyers, B. C., Nevo, E. & Michelmore, R. W. Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *The Plant cell* **16**, 2870-2894, doi:10.1105/tpc.104.025502 (2004).
43 Panchy, N., Lehti-Shiu, M. & Shiu, S.-H. Evolution of Gene Duplication in Plants. *Plant Physiology* **171**, 2294, doi:10.1104/pp.16.00523 (2016).

44 Zhang, L. *et al.* Rapid evolution of protein diversity by de novo origination in Oryza. *Nature Ecology & Evolution* **3**, 679-690, doi:10.1038/s41559-019-0822-5 (2019).
45 Dunning, L. T. *et al.* Lateral transfers of large DNA fragments spread functional genes among grasses. *Proceedings of the National Academy of Sciences* **116**, 4416-4425, doi:10.1073/pnas.1810031116 (2019).
46 Woodhouse, M. R. *et al.* Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. *PLoS biology* **8**, e1000409-e1000409 (2010).
47 Woodhouse, M. R., Pedersen, B. & Freeling, M. Transposed Genes in Arabidopsis Are Often Associated with Flanking Repeats. *PLOS Genetics* **6**, e1000949, doi:10.1371/journal.pgen.1000949 (2010).
48 Edger, P. P. *et al.* Origin and evolution of the octoploid strawberry genome. *Nat Genet* **51**, 541-547, doi:10.1038/s41588-019-0356-4 (2019).
49 Bird, K. A. *et al.* Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid Brassica napus. *bioRxiv*, 814491, doi:10.1101/814491 (2019).
50 Tang, H. *et al.* Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. *Genetics* **190**, 1563-1574, doi:10.1534/genetics.111.137349 (2012).
51 Cheng, F., Wu, J. & Wang, X. Genome triplication drove the diversification of Brassica plants. *Hortic Res* **1**, 14024-14024, doi:10.1038/hortres.2014.24 (2014).
52 Golicz, A. A. *Construction and analysis of the Brassica oleracea pangenome*, The University of Queensland, (2016).
53 Bird, K. A., VanBuren, R., Puzey, J. R. & Edger, P. P. The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytologist* **220**, 87-93, doi:10.1111/nph.15256 (2018).
54 Chalhoub, B. *et al.* Plant genetics. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* **345**, 950-953, doi:10.1126/science.1253435 (2014).
55 Samans, B., Chalhoub, B. & Snowdon, R. J. Surviving a genome collision: genomic signatures of allopolyploidization in the recent crop species Brassica napus. *The plant genome* **10** (2017).
56 Feldman, M., Levy, A. A., Fahima, T. & Korol, A. Genomic asymmetry in allopolyploid plants: wheat as a model. *Journal of experimental botany* **63**, 5045-5059 (2012).
57 Appels, R. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
58 Ramírez-González, R. *et al.* The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018).
59 Bardil, A., de Almeida, J. D., Combes, M. C., Lashermes, P. & Bertrand, B. Genomic expression dominance in the natural allopolyploid Coffea arabica is massively affected by growth temperature. *New Phytologist* **192**, 760-774 (2011).
60 Yoo, M., Szadkowski, E. & Wendel, J. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* **110**, 171-180 (2013).
61 Edger, P. P. *et al.* Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. *Plant Cell* **29**, 2150-2167, doi:10.1105/tpc.17.00010 (2017).
62 Kashkush, K., Feldman, M. & Levy, A. A. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* **160**, 1651-1659 (2002).
63 Hawkins, J. S., Proulx, S. R., Rapp, R. A. & Wendel, J. F. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proceedings of the National Academy of Sciences* **106**, 17811-17816 (2009).
64 Freeling, M. *et al.* Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome research* **18**, 1924-1937 (2008).

65 McClintock, B. Induction of instability at selected loci in maize. *Genetics* **38**, 579 (1953).
66 Ou, S. *et al.* Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline. *bioRxiv*, 657890 (2019).
67 Yan, H., Bombarely, A. & Li, S. DeepTE: a computational method for de novo classification of transposons with convolu-tional neural network. *bioRxiv* (2020).
68 da Cruz, M. H. P., Domingues, D. S., Saito, P. T. M., Paschoal, A. R. & Bugatti, P. H. TERL: Classification of Transposable Elements by Convolutional Neural Networks. *bioRxiv* (2020).
69 Van Oss, S. B. & Carvunis, A.-R. De novo gene birth. *PLOS Genetics* **15**, e1008160, doi:10.1371/journal.pgen.1008160 (2019).
70 Golicz, A. A., Bhalla, P. L. & Singh, M. B. lncRNAs in Plant and Animal Sexual Reproduction. *Trends in Plant Science* **23**, 195-205, doi:10.1016/j.tplants.2017.12.009 (2018).
71 Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *eLife* **3**, e03523, doi:10.7554/eLife.03523 (2014).
72 Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* **36**, 875-879, doi:10.1038/nbt.4227 (2018).
73 Rabbani, L., Mueller, J. & Weigel, D. An Algorithm to Build a Multi-genome Reference. *bioRxiv*, 2020.2004.2011.036871, doi:10.1101/2020.04.11.036871 (2020).
74 Jensen, S. E. *et al.* A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. *The Plant Genome* **n/a**, e20009, doi:10.1002/tpg2.20009 (2020).
75 Contreras-Moreira, B. *et al.* Analysis of Plant Pan-Genomes and Transcriptomes with GET_HOMOLOGUES-EST, a Clustering Solution for Sequences of the Same Species. *Front Plant Sci* **8**, 184, doi:10.3389/fpls.2017.00184 (2017).
76 Golicz, A. A., Bhalla, P. L. & Singh, M. B. MCRiceRepGP: a framework for the identification of genes associated with sexual reproduction in rice. *The Plant Journal* **96**, 188-202, doi:10.1111/tpj.14019 (2018).
77 Hassani-Pak, K. *et al.* Developing integrated crop knowledge networks to advance candidate gene discovery. *Applied & Translational Genomics* **11**, 18-26, doi:https://doi.org/10.1016/j.atg.2016.10.003 (2016).
78 Rodgers-Melnick, E., Vera, D. L., Bass, H. W. & Buckler, E. S. Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences* **113**, E3177-E3184 (2016).
79 Maistrenko, O. M. *et al.* Disentangling the impact of environmental and phylogenetic constraints on prokaryotic strain diversity. *bioRxiv*, 735696 (2019).
80 Lin, K. *et al.* Beyond genomic variation - comparison and functional annotation of three *Brassica rapa* genomes: a turnip, a rapid cycling and a Chinese cabbage. *BMC Genomics* **15**, 250, doi:10.1186/1471-2164-15-250 (2014).
81 Li, Y. H. *et al.* De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045-1052, doi:10.1038/nbt.2979 (2014).
82 Schatz, M. C. *et al.* Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. *Genome Biology* **15**, 506, doi:10.1186/s13059-014-0506-z (2014).
83 Hirsch, C. N. *et al.* Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* **26**, 121-135, doi:10.1105/tpc.113.119982 (2014).
84 Yao, W. *et al.* Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biology* **16**, 187, doi:10.1186/s13059-015-0757-3 (2015).
85 Golicz, A. A. *et al.* The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications* **7**, 13390, doi:10.1038/ncomms13390 (2016).
86 Pinosio, S. *et al.* Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Mol. Biol. Evol.* **33**, 2706-2719, doi:10.1093/molbev/msw161 (2016).

87 Gordon, S. P. *et al.* Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications* **8**, 2184, doi:10.1038/s41467-017-02292-8 (2017).
88 Zhou, P. *et al.* Exploring structural variation and gene family architecture with *de novo* assemblies of 15 *Medicago* genomes. *BMC Genomics* **18**, 261, doi:10.1186/s12864-017-3654-1 (2017).
89 Hurgobin, B. *et al.* Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnol. J.* **16**, 1265-1274, doi:10.1111/pbi.12867 (2018).
90 Ou, L. J. *et al.* Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence-absence variation analyses. *New Phytol.* **220**, 360-363, doi:DOI 10.1111/nph.15413 (2018).
91 Zhao, Q. *et al.* Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278-284, doi:10.1038/s41588-018-0041-z (2018).
92 Wang, W. S. *et al.* Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43-49, doi:10.1038/s41586-018-0063-9 (2018).
93 Yu, J. Y. *et al.* Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnol. J.* **17**, 881-892, doi:10.1111/pbi.13022 (2019).
94 Hubner, S. *et al.* Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat Plants* **5**, 54-62, doi:10.1038/s41477-018-0329-0 (2019).
95 Gao, L. *et al.* The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044-1051, doi:10.1038/s41588-019-0410-2 (2019).
96 Song, J. M. *et al.* Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat Plants* **6**, 34-45, doi:10.1038/s41477-019-0577-7 (2020).
97 Trouern-Trend, A. J. *et al.* Comparative genomics of six Juglans species reveals disease-associated gene family contractions. *Plant J.* **102**, 410-423, doi:10.1111/tpj.14630 (2020).
98 Liu, Y. *et al.* Pan-genome of wild and cultivated soybeans. *Cell*, doi:10.1016/j.cell.2020.05.023 (2020).

Correspondence

Correspondence should be addressed to Professor David Edwards,

dave.edwards@uwa.edu.au

Acknowledgments

This research was supported by the Australian Government through the Australian Research Council's Linkage Projects funding scheme (project LP140100537, LP160100030). PB acknowledges support of the Forrest Research Foundation. AS was supported by NSF grant IOS-1822330.

Author contributions

PB, AG, AS, JB, and DE wrote and edited the manuscript together.

Competing Interests Statement

The authors declare no competing interests.

Figure Legends

Figure 1: Comparison of pangenome approaches. a) Alignment of reads from multiple samples to a reference is followed by assembly of unaligned reads into novel contigs. By adding these novel contigs to the original reference sequence, a pangenome reference can be constructed. Dispensable regions are determined based on mapping all reads back to the pangenome. b) De novo assembly of the genomes of multiple accessions allows whole genome alignment approaches to identify dispensable genomic regions. c) A pangenome graph can be constructed from whole genome alignments or by *de novo* graph assembly, and efficiently stores variant information of dispensable regions as unique paths through the graph.

Figure 2: Number of times the terms ‘pangenome’ or ‘pan-genome’ are mentioned in Europe Pubmed Central from the first mention in Tettelin et al. 2008 to 2019.

Figure 3: Four sources for novel genes. A) Whole Genome Duplication, in which a genome is completely duplicated leading to a duplicated set of genes. After the WGD event, genes are slowly lost. B) Tandem duplication, in which a region is locally duplicated leading to neighbouring identical gene copies. C) transposable element (TE) mediated insertion, in which a transposon carries copies of genes into other regions of the genome. D) de novo gene birth, in which novel open reading frames are created from other gene fragments or non-coding regions. E) segmental duplication, where an entire region is duplicated in the genome. F) Introgression, where a genomic region from the same or closely related species is introgressed. G) Horizontal gene transfer, where a genomic region from a different species or even different kingdom is introgressed.