Molecular Plant
Opinion

CellPress
Partner Journal

Toward haplotype studies in polyploid plants to assist breeding

THE IMPORTANCE OF POLYPLOIDS

Polyploids are typically classified as either autopolyploids or allopolyploids (Figure 1). Autopolyploids result from whole-genome duplication within the same species, while allopolyploids derive from the hybridization of different species followed by chromosome doubling. Taxonomically, plant allopolyploids are thought to be the most common polyploids, although autopolyploid plants and allopolyploid plants might be at parity in numbers (Barker et al., 2016). During speciation, polyploidization allows plants to adapt to different environments (Soltis et al., 2009). Mutation and hybridization increase the heterozygosity of the genome, while genome rearrangements during polyploidization lead to the formation of new chromosomes and new chromosome rearrangements, which complicate polyploid genomes and the following studies.

Many crops are recent polyploids, including wheat, potato, banana, peanut, canola, and cotton. They are key staple food crops that provide substantial amounts of nutrition to humans (FAO, 2020). Improving the productivity of these crops and their adaptation to different environments is crucial. However, current breeding programs are facing bottlenecks to continue producing desired polyploid varieties. One of the reasons might be that over long periods of selective breeding by humans, some useful genes in polyploids have been unintentionally discarded to meet previous breeding purposes and those genes are difficult to retrieve (Smýkal et al., 2018). Another reason might be that the complexity and repetitiveness of polyploid genomes make it more difficult to associate desirable phenotypes with the underlying genetic variants, because the variations in polyploid genomes, such as segmental duplications and homoeologous exchanges, may reduce the accuracy of bioinformatics tools for genome assembly and genotype discovery. One more reason might be that the limitations of previous technologies, methods and relatively high cost of orthogonal data production have made it challenging to comprehensively characterize polyploid genomes, particularly for autopolyploid genomes and heterozygous genomes broken up by homozygous regions (Zhang et al., 2019). Therefore, developing new solutions, particularly identifying useful genetic markers using new technology and computational methods, is important to polyploid plant breeding.

THE PITFALLS IN PREVIOUS SNP-GWAS ON POLYPLOIDS

Although tremendous efforts have been applied to the study and breeding of polyploids using different genetic markers, most recent work has examined the association between individual single-nucleotide polymorphisms (SNPs) with desired traits, either at the whole-genome level or at the level of an individual chromosome or locus, using genome-wide association studies (GWAS) (Brachi et al., 2011). Individual SNPs may not be the causal polymorphisms, and the SNPs investigated in GWAS are usually exclusively biallelic. Such practices are useful for many phenotype-genotype association studies. However, they may overlook multiallelic SNPs or a cluster of SNPs or genes (haplotypes) with *cis* or *trans* variations in polyploid alleles that may be responsible for complex and quantitative traits. Moreover, as a result of genome rearrangements, for example, homoeologous exchanges in allopolyploids, some of the sequence segments in polyploid genomes may be duplicated in different chromosome regions. This could confuse read aligners and make SNP calling less accurate, particularly when relying on short sequence reads. Thus, accurate SNP identification in polyploid genomes is more challenging.

THE IMPORTANCE OF POLYPLOID HAPLOTYPE STUDIES

Recently, several studies have unveiled that some important agronomic traits are governed by multiple haplotypes, such as the haplotypes surrounding the productivity-related *TaGW2-A* gene and the haplotypes within the insect resistance *Sm1* locus in wheat (Brinton et al., 2020; Todesco et al., 2020; Walkowiak et al., 2020). Therefore, resolving different haplotypes or haplotype blocks in polyploids that have long been neglected would be valuable for polyploid crop improvement. By using methods such as linkage-disequilibrium-based GWASs, researchers can determine the association between individual haplotype alleles with the desired agronomic trait and then select these alleles for breeding.

In addition, studying haplotypes in polyploid plant genomes can shed light on the evolution and domestication of crop species, as well as reveal their breeding history. A complete survey of standing variation and its origin can guide breeders to develop novel variations for breeding. Resolving haplotypes in polyploids can not only reveal patterns of genetic heredity from parents to offspring but also indicate gene expression or protein function dominance caused by different haplotypes due to *cis* or *trans* variations at homologous loci. Such allele-specific analyses are of particular relevance for the dissection of heterosis and subgenome dominance in polyploids. Studies of haplotypes in polyploids can also facilitate the understanding of genome

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

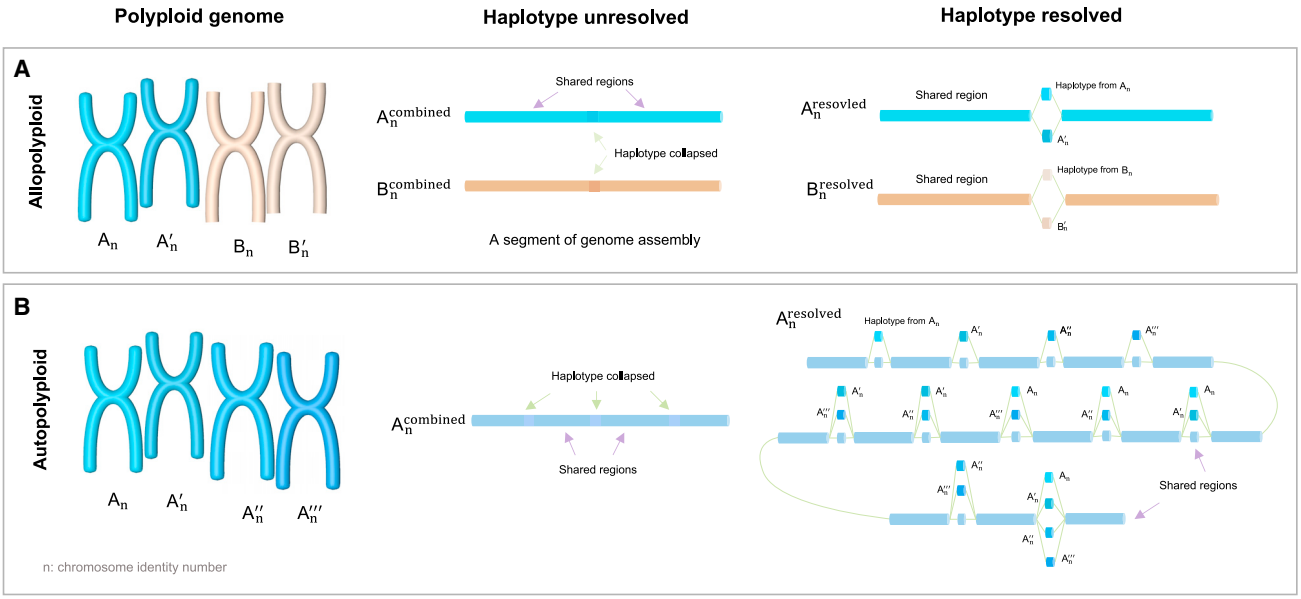

Figure 1. Illustration of unresolved and resolved haplotypes in polyploid genomes.
An allotetraploid genome (**A**) and an autotetraploid genome (**B**) are shown. In both cases, the currently widely adopted approaches (middle column) provide only a combined monoploid representation of homologous chromosome collapsing haplotypes. Consequently, this could create falsely linked genetic variants from different haplotypes. Haplotype-aware genome assembly or haplotype resolving at each locus (right column) could mitigate these issues by correctly capturing alternative haplotypes and assigning them to their progenitor chromosome. Using haplotype-aware genome assemblies or efficiently resolving each haplotype in polyploid genomes, it would be easier to mine different haplotypes in the test samples and associate phenotypes with genotypes, therefore benefiting agronomic gene exploration and advanced breeding.

recombination or sequence exchanges by comparing the genetic information with that of parents or ancestors.

In some cases, the progenitors or parents of some polyploids are unknown. By constructing a haplotype-aware reference panel in a population, it is possible to resolve haplotypes in the studied sample by imputation (statistically inferring genotypes by using known haplotypes in a population) (Chen et al., 2021), thus increasing the accuracy of phenotype-genotype association analyses. Furthermore, genetic exchanges between genomes following polyploidization can also be studied.

RESOLVING POLYPLOID HAPLOTYPES

In contrast to diploid organisms, which have two alleles at the same locus, polyploids have multiple alleles, which complicates the identification of different haplotypes on the alleles (Schrinner et al., 2020). To better resolve haplotypes in polyploid genomes, it is possible to use efficient chromosome set separation before genotyping. After decades of attempts, researchers have successfully separated homologous chromosome sets in allopolyploids using approaches relying on DNA sequencing and novel bioinformatics tools (Kyriakidou et al., 2018) that benefit from the diploid-like nature of allopolyploids. Nevertheless, efficient sequence isolation in autopolyploid genomes is still challenging owing to the high similarity between each homologous chromosome set and the possible meiotic recombinants (Figure 1B). While allopolyploids such as teff show diverged subgenomes with coding sequence similarities of 93.9% (VanBuren et al., 2020), autoploids may show much higher similarities, as illustrated by the 99% subgenome sequence similarity between allelic chromosome pairs in cultivated alfalfa (Chen et al., 2020). Consequently, most autopolyploid assemblies are presented as monoploid sequences, which mosaic, collapse, or eliminate different haplotypes or homologous copies (Zhang et al., 2019). This oversimplification of the true genome sequences could artificially conceal allelic variations in polyploid genomes and impede downstream analyses of genotyping and the associated cellular and biological functions (Zhang et al., 2020).

Owing to the limitations of previous technologies and methods, most polyploid plant genomes are not yet haplotype resolved, although it is relatively easier to do so in the diploid-like allopolyploids (Figure 1A). During genome assembly, many allelic variants or heterozygous regions are generally initially resolved in the form of graph bubbles. However, in the final assembly, these assembly bubbles have not been efficiently anchored or assigned to a progenitor chromosome, leaving the assembly haplotype unaware.

It is feasible to resolve haplotypes in each polyploid allele using, for instance, SNP arrays or whole-genome sequencing (Schrinner et al., 2020; Zhou et al., 2020). However, owing to the fixed set of SNPs assayed with a genotyping array, the limited read length in short-read sequencing, and the relatively high raw error rates (5%–15%) in commonly used single-molecule long-read sequencing, haplotype resolving in polyploid plant genomes is still limited.

With recent advances in PacBio HiFi sequencing and Oxford Nanopore R10.3 sequencing, and together with Oxford Nanopore ultra-long-read sequencing, high-throughput chromosome

conformation capture sequencing, single-cell DNA template strand sequencing,and improved optical mapping, accurately resolving a haplotype in polyploid plant genomes is becoming practical either by mapping or by *de novo* assembly. Given the technology and method (see Supplemental Table 1) breakthrough, it is also possible to move from the current linear genome assembly to a haplotype-resolved polyploid genome assembly or a graph genome assembly to clearly capture different haplotypes at each locus (Figure 1B). While the cost of long-read sequencing is currently relatively high, it is expected to decline.

CONCLUDING REMARKS

The resolution of haplotypes in polyploid genomes can uncover alternative genotypes that have been missed in previous studies. By associating these hidden haplotypes with phenotypes, we can gain insights into the evolution, domestication, and breeding of polyploids.

The previous use of DNA sequencing for resolving haplotypes was limited by the read length in short-read sequencing and the high error rate in commonly used long-read sequencing approaches. The use of monoploid assemblies could have introduced bias in variant calling and GWASs. Large and complex haplotype blocks could also be collapsed in monoploid assemblies. Taken together, this has impeded the exploration and use of different haplotypes in polyploid genomes.

With recent advances in technologies and methods, more efforts are needed to develop novel algorithms, such as machine learning algorithms to resolve the alternative haplotypes in polyploids. Importantly, tools used in diploid haplotype studies, such as hifiasm (Cheng et al., 2021), are expected to support polyploid genomes in the future. User-friendly genome visualization tools, for instance, using a graph to display genomes, are also expected to show the resolved haplotypes at each locus in one genome or a pangenome. Assembly, phenotype, and environmental variation-integrated databases are needed as well to document important haplotypes/genes associated with desired agronomic traits, thereby providing valuable resources for polyploid plant breeding. With the advances in genome editing, haplotypes could also be edited in the polyploid genome to activate preferred genes or silence unwanted genes and thus assist the production of elite crop varieties.

Haplotypes and their potential functions in polyploids are important but remain largely unexplored. Greater attention on this topic can allow us to gain a deeper understanding of polyploid evolution and domestication and ultimately help us to breed better polyploid crops.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported by the Hong Kong Research Grants Council Area of Excellence Scheme (AoE/M-403/16) and Collaborative Research Fund (C4057-18EF), CUHK Group Research Scheme 3110135, and the Innovation and Technology Commission, Hong Kong Special Administrative Region Government to the State Key Laboratory of Agrobiotechnology (CUHK). A.S. was supported by NSF grant IOS-1822330.

ACKNOWLEDGMENTS

Any opinions, findings, conclusions, or recommendations expressed in this publication do not reflect the views of the government of the Hong Kong Special Administrative Region or the Innovation and Technology Commission. No conflict of interest is declared.

Yuxuan Yuan[1], Armin Scheben[2], David Edwards[3] and Ting-Fung Chan[1,]*

[1]School of Life Sciences and State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[3]School of Biological Sciences and Institute of Agriculture, The University of Western Australia, Perth, Australia
***Correspondence: Ting-Fung Chan (tf.chan@cuhk.edu.hk)**
https://doi.org/10.1016/j.molp.2021.11.004

REFERENCES

Barker, M.S., Arrigo, N., Baniaga, A.E., Li, Z., and Levin, D.A. (2016). On the relative abundance of autopolyploids and allopolyploids. New Phytol. **210**:391–398. https://doi.org/10.1111/nph.13698.

Brachi, B., Morris, G.P., and Borevitz, J.O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol. **12**:232. https://doi.org/10.1186/gb-2011-12-10-232.

Brinton, J., Ramirez-Gonzalez, R.H., Simmonds, J., Wingen, L., Orford, S., Griffiths, S., Wheat Genome, P., Haberer, G., Spannagl, M., Walkowiak, S., et al. (2020). A haplotype-led approach to increase the precision of wheat breeding. Commun. Biol. **3**:712. https://doi.org/10.1038/s42003-020-01413-2.

Chen, H., Zeng, Y., Yang, Y., Huang, L., Tang, B., Zhang, H., Hao, F., Liu, W., Li, Y., Liu, Y., et al. (2020). Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat. Commun. **11**:2494. https://doi.org/10.1038/s41467-020-16338-x.

Chen, M., Fan, W., Ji, F., Hua, H., Liu, J., Yan, M., Ma, Q., Fan, J., Wang, Q., Zhang, S., et al. (2021). Genome-wide identification of agronomically important genes in outcrossing crops using OutcrossSeq. Mol. Plant **14**:556–570. https://doi.org/10.1016/j.molp.2021.01.003.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods **18**:170–175. https://doi.org/10.1038/s41592-020-01056-5.

FAO. (2020). Crop Prospects and Food Situation - Quarterly Global Report No. 2 (Rome, Italy: FAO).

Kyriakidou, M., Tai, H.H., Anglin, N.L., Ellis, D., and Stromvik, M.V. (2018). Current strategies of polyploid plant genome sequence assembly. Front Plant Sci. **9**:1660. https://doi.org/10.3389/fpls.2018.01660.

Schrinner, S.D., Mari, R.S., Ebler, J., Rautiainen, M., Seillier, L., Reimer, J.J., Usadel, B., Marschall, T., and Klau, G.W. (2020). Haplotype threading: accurate polyploid phasing from long reads. Genome Biol. **21**:252. https://doi.org/10.1186/s13059-020-02158-1.

Smýkal, P., Nelson, M.N., Berger, J.D., and Von Wettberg, E.J.B. (2018). The impact of genetic changes during crop domestication. Agronomy **8**:119.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Depamphilis, C.W., Wall, P.K., and Soltis, P.S. (2009). Polyploidy and angiosperm diversification. Am. J. Bot. **96**:336–348. https://doi.org/10.3732/ajb.0800079.

Todesco, M., Owens, G.L., Bercovich, N., Legare, J.S., Soudi, S., Burge, D.O., Huang, K., Ostevik, K.L., Drummond, E.B.M., Imerovski, I., et al. (2020). Massive haplotypes underlie ecotypic

differentiation in sunflowers. Nature **584**:602–607. https://doi.org/10.1038/s41586-020-2467-6.

VanBuren, R., Man Wai, C., Wang, X., Pardo, J., Yocca, A.E., Wang, H., Chaluvadi, S.R., Han, G., Bryant, D., Edger, P.P., et al. (2020). Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. Nat. Commun. **11**:884. https://doi.org/10.1038/s41467-020-14724-z.

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M.T., Brinton, J., Ramirez-Gonzalez, R.H., Kolodziej, M.C., Delorean, E., Thambugala, D., et al. (2020). Multiple wheat genomes reveal global variation in modern breeding. Nature https://doi.org/10.1038/s41586-020-2961-x.

Zhang, X., Zhang, S., Zhao, Q., Ming, R., and Tang, H. (2019). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat. Plants **5**:833–845. https://doi.org/10.1038/s41477-019-0487-8.

Zhang, X., Wu, R., Wang, Y., Yu, J., and Tang, H. (2020). Unzipping haplotypes in diploid and polyploid genomes. Comput. Struct. Biotechnol. J. **18**:66–72. https://doi.org/10.1016/j.csbj.2019.11.011.

Zhou, C., Olukolu, B., Gemenet, D.C., Wu, S., Gruneberg, W., Cao, M.D., Fei, Z., Zeng, Z.B., George, A.W., Khan, A., et al. (2020). Assembly of whole-chromosome pseudomolecules for polyploid plant genomes using outbred mapping populations. Nat. Genet. **52**:1256–1264. https://doi.org/10.1038/s41588-020-00717-7.