


Features and applications of haplotypes in crop breeding

Javaid Akhter Bhat¹, Deyue Yu¹, Abhishek Bohra², Showkat Ahmad Ganie ³✉ & Rajeev K. Varshney ^{4,5}✉

Climate change with altered pest-disease dynamics and rising abiotic stresses threatens resource-constrained agricultural production systems worldwide. Genomics-assisted breeding (GAB) approaches have greatly contributed to enhancing crop breeding efficiency and delivering better varieties. Fast-growing capacity and affordability of DNA sequencing has motivated large-scale germplasm sequencing projects, thus opening exciting avenues for mining haplotypes for breeding applications. This review article highlights ways to mine haplotypes and apply them for complex trait dissection and in GAB approaches including haplotype-GWAS, haplotype-based breeding, haplotype-assisted genomic selection. Improvement strategies that efficiently deploy superior haplotypes to hasten breeding progress will be key to safeguarding global food security.

Crop plants are subjected to a variety of biotic and abiotic stresses that impair normal crop growth and cause substantial losses in crop yields worldwide^{1,2}. Amid these stresses, developing climate smart and nutritious crop varieties that remain vital to securing food security of the incessantly growing human population, presents a daunting challenge to the agricultural scientists worldwide. Although conventional breeding has made great success in the development of high-yielding crop varieties³, it is important to accelerate the pace of crop improvement programmes especially for the complex traits such as yield under stress conditions. In this regard, the genomics-assisted breeding (GAB) by implementing genomics tools in breeding was proposed by Varshney et al.⁴. This approach has delivered several high-yielding, stress-tolerant and better nutrition varieties^{5,6}. For instance, the low-throughput sequence-based markers, such as simple sequence repeats (SSRs), were extensively used in the molecular breeding programmes; however, these marker systems have limitations such as low density across the genome, low coverage, expensiveness. Application of these second-generation DNA marker systems resulted in poor resolution of gene mapping and relatively low efficiency of plant selections and breeding^{7,8}. Fortunately, recent advances in the next generation sequencing (NGS) and the genotyping platforms have considerably alleviated this bottleneck in crop breeding. These NGS-based platforms have provided remarkable marker-density and coverage at reduced cost⁹, and are now commercially available for both model and non-model crop species^{10,11}. These high-throughput platforms make hundreds of millions of DNA polymorphisms accessible for use in genetic and genomics research^{12,13}, and their application in crop breeding has considerably increased the gene mapping resolution and prediction accuracy in genomic selection (GS)^{14,15}. Majority of the economically important crop traits, such as yield, quality and stress tolerance, are of complex quantitative nature, which are influenced by several small effect QTL/genes and manifest substantial genotype x environment (G x E) interactions¹⁶. Although efforts

¹National Center for Soybean Improvement, State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China. ²Crop Improvement Division, ICAR- Indian Institute of Pulses Research (ICAR- IIPR), Kanpur, India. ³Department of Biotechnology, Visva-Bharati, Santiniketan 731235 WB, India. ⁴Center of Excellence in Genomics & Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad 502324, India. ⁵State Agricultural Biotechnology Centre, Centre for Crop & Food Research Innovation, Food Futures Institute, Murdoch University, Murdoch, WA, Australia. ✉email: showkatmanzoorforever@gmail.com; rajeev.varshney@murdoch.edu.au

to understand the complex genetic makeup of these agriculturally relevant traits have been successful in the identification of major-effect genomic regions, conventional experimental populations faced the problem of limited genetic diversity, low resolution and limited recombination events^{17,18}. Hence, the genome-wide association study (GWAS) has emerged as a powerful tool for dissecting complex quantitative traits in crop plants with enhanced resolution and allelic richness^{19,20}. Furthermore, due to the availability of cost-effective and high-density genotyping platforms, it has been possible now to screen larger breeding populations for estimating and using the breeding value in crop improvement programmes by using GS, another breeding approach²¹.

In recent years, the NGS-based genotyping methods such as genotyping-by-sequencing, restriction site-associated DNA sequencing, whole-genome resequencing as well as fixed SNP arrays have greatly facilitated genotyping of large germplasm collections for GWAS and GS analyses^{8,22}. However, the major limitations for the use of SNPs in these analyses include their biallelic nature, the presence of rare alleles, and abundant levels of linkage drag^{16,23}. Therefore, the candidate genomic loci identified by GWAS often do not represent the causative locus; but correspond to the loci that are in linkage drag with a gene or a regulatory element, eventually affecting the trait of interest^{24,25}. In this regard, an effective approach to overcome the limitations of SNPs and increase the resolution of candidate genomic regions is to consider haplotypes for genome-wide analyses²⁶. Haplotype is a specific combination of jointly inherited nucleotides or DNA markers from polymorphic sites in the same chromosomal segment^{27,28}.

In the present review, we discuss the potential and need of haplotypes in the crop breeding for the development of improved varieties. We have also compared the efficiency of haplotype- and individual SNP-based markers in the GWAS and GS analyses. Besides, the challenges associated with the use of haplotypes in crop breeding at the commercial level are also addressed. We conclude by highlighting the scope of haplotypes in the future crop breeding programs.

Crop improvement: conventional breeding to genomics-assisted breeding. Development of improved crop varieties for food, feed and industrial purposes can be accomplished mainly by plant breeding²⁹. The science of plant breeding has evolved from conventional to present day GAB^{6,30}. In the last century, tremendous efforts have been made by plant breeders across the globe to develop improved varieties in different crop species by using the conventional breeding approaches^{31–45}. It is estimated that the undernourished proportion of the human population has been reduced from 40% in the 1960s to <11% now, which is principally attributable to the improved high-yielding and stress-tolerant crop varieties produced mainly through conventional breeding⁴⁴. The conventional plant breeding for crop yield enhancement progressed consistently over time. The high-yielding varieties/hybrids were mostly responsible for this increase in both area and productivity, and the large-scale adoption of these varieties/hybrids provides strong evidence for contributions by plant breeding innovations over the last century.

In recent years, the plant breeding community has recognized the need of introducing genetic variability in breeding programs to enhance the genetic base of elite gene pool, enhancing precision and efficiency in selection and reducing the breeding cycle^{4,6,46}. In this context, the GAB approach proposed by Varshney et al.⁴ outlined the use of genomics tools and technologies to identify markers, candidate genes associated with target traits and integration of genomics approaches in breeding.

Several GAB approaches including marker-assisted backcrossing (MABC), marker-assisted selection (MAS), marker-assisted recurrent selection (MARS) and advanced backcross QTL (AB-QTL) were suggested for crop improvement. In recent years, GS approach has also been added to GAB portfolio^{6,21}. For MAS, the first step is the identification of molecular markers that are strongly associated with genomic regions/quantitative trait loci (QTLs) regulating the traits of interest. Eventually, these QTLs, either individually or in multiple numbers, can be pyramided into elite breeding material through MABC. Some success stories of MABC include the introgression of a 'QTL-hotspot' into elite chickpea varieties for improved yield under drought conditions^{47,48}, improving the yield and stress tolerance of mega rice variety IR64 (Developed by IRRI, IR 64 was released in Philippines in 1987. The rice variety registered a widespread acceptance owing to its multiple beneficial traits including better cooking quality, earliness, disease resistance and high yield)^{49,50}, transferring QTLs (*qDTY2.2* and *qDTY4.1*) into IR64 for reproductive stage drought tolerance^{51,52}, and the improvement of different yield and stress-related traits in several major crop species^{6,53–55}. Despite the aforementioned utilities of MABC, it is efficient only for the major-effect QTLs, while most of the genetic variations for yield, quality and stress tolerance traits in crop plants are governed by a large number of minor QTLs. Alternatively, the frequency of many beneficial alleles can be increased in a given population through the MARS scheme. Unlike MABC, the MARS has been applied for improving a breeding population with respect to QTLs exerting smaller effects on the phenotype. MARS has been successful in improving drought tolerance in multiple crop species viz., maize, soybean, sunflower, wheat, sorghum, and rice^{56–60}. To capture minor effect QTLs scattered throughout the genome, the plant breeding community has recently started to use GS approach. GS estimates the genetic worth of an individual based on the large set of marker information distributed across the whole genome, rather than a few markers as in the case of MAS²¹. In this approach, a prediction model based on the genotypic and phenotypic data of training population (TP) is developed and then genomic estimated breeding values (GEBVs) for the individuals of breeding population (BP) are computed from their genome-wide marker profiles⁶¹. The GEBVs allow one to predict individuals that will perform better and are suitable either as a parent for the next breeding cycle or can directly enter into the variety release pipeline²¹. Unlike MAS, GS does not necessarily require a prior knowledge of significant marker-trait associations⁶². However, inclusion of the significant set of markers, such as resulting from GWAS, into GS models has been found to improve prediction accuracies⁶³. GS has started gaining profound interest in plant breeding, with the recent studies establishing its superiority over other selection methods^{64–70}. With the availability of a range of cost-effective genotyping platforms and advances in the development of prediction models, GS is expected to be a routine breeding approach, like MABC/MAS in crop improvement programmes.

Features of haplotypes

Defining haplotypes: harnessing the wealth of whole-genome sequencing data. Haplotype is a combination of alleles for different polymorphisms (such as SNPs, insertions/deletions and other markers or variants) present on the same chromosome, which are inherited together with minimum chance of contemporary recombination^{71,72}. Any individual has two haplotypes for a given stretch of chromosomal DNA; while at the population level, many haplotypes can be found for the same stretch⁷³. In other words, a haplotype is defined as a set of nearby genomic

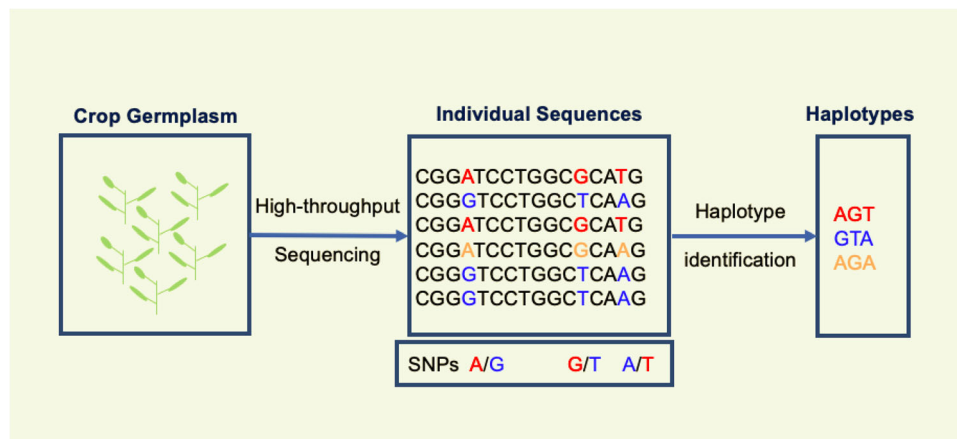


Fig. 1 Formation and development of haplotypes from haploid sequences. Resequencing of the crop germplasm is done to identify the polymorphic SNPs to be subsequently used in the development of haplotypes.

structural variations, such as polymorphic SNPs, with a strong linkage disequilibrium (LD) between them⁷⁴. As shown in Fig. 1, two or more polymorphic SNPs of the haploid sequences inherited together as a unit constitute a haplotype⁷¹. The haplotypes are defined/assigned in three principal ways: (a) by using the haplotype diversity in a given chromosomal segment, (b) by using the pairwise LD between the jointly inherited markers that show lack of evidence for historical recombination, it is measured by r^2 (measure of LD)^{75,76} and (c) by grouping of SNPs through sliding-windows of fixed or variable length⁷⁷. Evidence suggests that the LD-based approaches are more efficient for defining the haplotypes in the genomic/chromosome regions^{26,74}. This is because (a) historical recombination identification is the direct focus in a particular population through the haplotype detection, (b) visualization of the LD coefficients is very easy, (c) for diploid data with unknown haplotype phase, it is applicable. The LD in the given population is determined by many factors such as mode of pollination, population size and structure, mutation rate, genetic drift, recombination frequency, and the type of selection on a given chromosomal fragment⁷⁸.

During the evolution of the important crop species such as rice, maize, wheat, sorghum, cassava and rapeseed, the selection of genes/alleles regulating desirable phenotype for the trait of interest is the major factor responsible for the formation of signatures of selection²⁶. The signatures of selection (also known as conserved haplotype blocks and selective sweeps) possess multiple genes, which are regulated together by many regulatory genes. The correlation among different traits as reflected from the selection signatures is either due to the true linkage among the genes or resulting from the pleiotropic effect of the same genes^{34,79}. Therefore, the crop breeders should preferably target these genomic regions to elucidate their effect on the traits of interest. Besides, the integration of genomics to identify the recombinants produced by crossing of contrasting parents will greatly assist in resolving the complexity of quantitative traits. This will enhance the efficiency to improve the specific traits in modern varieties for their better adaptation to extreme environments⁸⁰.

Due to the availability of sequencing data from large number of individuals for a given crop species it has been easier to define the haplotype. By using the whole genome sequencing data, Bevan et al.⁸¹ defined the concept of the haplotype assembly. Together with the phenotyping data of germplasm/breeding lines, it is possible to assess and validate phenotypic effects of the 'component' haplotypes. Based on this premise, and by using large-scale whole-genome resequencing datasets in combination

with *haplo-pheno* analysis, Abbai et al.⁸² identified useful haplotypes for future breeding in rice and Sinha et al.⁴⁶ followed the similar approach in pigeonpea. High-density SNP data generated from multiple genotypes via NGS-based or array-based approaches have been used for the development of haplotypes in many plant species. These haplotypes have also been used for various applications in research and breeding in different crop species (see details in Tables 1, 2).

Third-generation sequencing: alleviating the bottlenecks in haplotype identification. The long-term goal of genetics is to elucidate the effect of DNA sequence variations on the plant traits, and how these variations have led to the evolution of different populations and species^{83,84}. In genetics, linkage is a core concept on which molecular mapping of genetic determinants relies. For example, in the linkage or association mapping, the individual genetic markers/variants are used to determine their association(s) with the trait(s) of interest, instead of pinpointing the causal mutation³. The trait-associated DNA markers are then used as surrogates for the selection of the desirable phenotypes⁵. As we mentioned in the previous section, fast-tracking the process of targeted trait improvement will require a paradigm shift from individual SNP markers to haplotypes. The information on haplotypes regulating the important phenotypes is currently limited in the genetic studies⁸⁵, which prevents the accurate determination of ancestry reconstruction, rearrangements of chromosomes, allele-specific expression, and detection of selective sweeps^{86,87}.

However, the availability of the high-throughput sequencing platforms has made a tremendous impact on the identification of haplotypes and their application in the genetic studies. Although, the second-generation sequencing techniques produce short reads of 150 bp, these small reads normally do not possess more than a single variant⁸⁸. Hence, the haplotypes are constructed indirectly from this data and this needs specific statistical inferences from population genotyping data, which in turn increases the time and cost for the haplotype construction^{88,89}. In contrast, third-generation sequencing (TGS), such as the Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), produce long reads from which the haplotypes can be directly constructed⁸⁸. In comparison to the second-generation sequencing methods, analysis of DNA molecules can be performed directly via long-read sequencing platforms⁹⁰. However, the 'phasing' is used for some adjustment of the long-read sequencing data to increase the efficiency for haplotype identification. Construction of the haplotypes from the sequence data through

Table 1 The use of haplotype markers in genome-wide association mapping (GWAS) analyses in different crop species.

Crop Species	Trait	Population size	Haplotype markers	Haplotype-trait associations	PVE (%)	Reference
Soybean	100-seed weight; plant height; seed yield	169	941	87	9.14-15.83	134
Soybean	Agronomic and yield-related traits	296	-	10	>10.0	153
Wheat	Heading date; plant height; 1000-grain weight; grain number per spike; fruiting efficiency at harvest	102	4516	97	-	121
Wheat	Grain yield; days to heading; plant height	6461	519	36	2.2-5.6	154
Barley	Deoxynivalenol content in kernels; heading time; days to maturity; grain yield; plant height; specific weight; 1000-kernel weight	277	14,400	-	2.0-14.0	135
Barley	Yield and quality-related traits	106	2770	23	>10.0	131
Rice	Grain shape	372	-	30	-	155
Rice	Agronomic traits	414	15,275	-	-	109
Maize	Agronomic and reproductive traits	322	53,403	44	5.6-17.0	156
Maize	Total plant height; ear height; ear height/plant height	183	7,831	40	7.0-22.0	74
Maize	Agronomic and reproductive traits	>1000	154,104	-	>10.0%	157
Oat	Heading date	4657	164741	184	-	158
Rapeseed	Days to flowering; seed glucosinolate content	950	-	15	-	152

haplotype estimation is known as phasing; which is very important to elucidate the sequence-specific variations such as the effect of methylation, specific expression of alleles and compound heterozygosity⁹¹. Fixing of higher error rate (~10%) in the long-read sequencing technologies compared to short-read sequencing methods (NGS methods) needs specific bioinformatics-mediated adjustments⁹². In this regard, many different phasing methods enabling haplotype construction/reconstruction from long-read sequencing data have been recently developed, such as reference-based phasing (molecular haplotyping, single-cell phasing, and polyploid phasing), *de novo* genome assembly (such as diploid and polyploid assembly) and strain-resolved metagenome assembly (de novo re-assembly, single nucleotide variant-based assembly, read and contig binning)⁷². Combination of these haplotype analysis methods with various computational tools such as WhatsHap, HapCut2, HapTree, WhatsHap-polyphase, Falcon phase, Hifiasm, SDip, POLYTE, DESMAN, MetaMaps, and ProxiMeta, has greatly enhanced the efficiency and precision in the identification of *de novo* and rare variants from the long-read sequencing data⁷². Therefore, integrating the various phasing and bioinformatics tools with the long-read sequencing technologies has allowed us to fully exploit the potential of these sequencing approaches in haplotype construction⁹¹. For example, Ammar et al.⁷³ showed that MinION nanopore sequencer efficiently resolved the variants/haplotypes of *HLA-A*, *HLA-B* and *CYP2D6* genes by producing the long reads without even using the statistical phasing. Similarly, Zhang et al.⁹³ also demonstrated the higher accuracy of Nanopore sequencing in the identification of haplotypes across the genomes. Besides, recent advances in the PacBio's HiFi technology have allowed to produce long reads in the range of 15-20 Kb, with an error rate comparable to the second-generation sequencing i.e., more than 99% accuracy was achieved⁹⁴. These advancements have allowed reconstruction of the previously impossible near-complete human haplotypes that include microsatellites, repetitive elements, and other complex structural variations⁹⁵. Moreover, Sun et al.⁹⁶ used the PacBio HiFi reads (30x per haplotype) and hifiasm to produce the assembly of the autotetraploid genome of potato. This was the first study demonstrating the haplotype-resolved assembly of potato crop. Through single-cell genotyping and high-quality long-read sequencing of the tetraploid plants, the authors successfully reconstructed all four haplotypes showing considerably higher diversity among themselves. This haplotype diversity is significantly higher than the diversity commonly found within a given species. This evidenced that successful haplotype reconstruction in the polyploid species has a huge impact on breeding these crops in the future⁹⁶. Recent research demonstrates the enormous potential of the TGS in resolving the accuracy issues in the haplotype identification, thereby increasing the scope of haplotypes for genetic studies in both animals and plants⁷². Hence, the TGS platforms offer promising alternative to obtain haplotype-related information from the genomes, and future affordability of these sequencing platforms will have a profound impact on plant research and breeding.

Haplotagging: A novel sequencing strategy for rapid discovery of haplotypes. Recently, a simple, rapid and promising technique for linked-read (LR) sequencing (called 'haplotagging') has emerged^{97,98}. In this technique, molecular barcoding of long DNA molecules is carried out prior to sequencing, which in turn retains the long-range information by preserving the linked variants⁸⁵. The shared barcode is then used to link the individual short reads for constructing the original haplotype⁹⁸. However, currently the commercial utilization of haplotagging in the genetic studies is prevented by certain factors, which include the requirement of

Table 2 The use of haplotype markers in genomic selection in different crop species.

Crop Species	Trait	Training Population size	Haplotype markers	GS prediction accuracy	Reference
Bluegum	Traits related to wood quality and tree growth	646	~3000	0.58	105
Soybean	Plant height & grain yield per plant	235	357	>0.80 & >0.45	159
Sorghum	Agronomic and yield-related traits	207	1,974	0.57-0.73	160
Wheat	Yield, test weight, and protein content	383	1400	>0.40	151
Wheat	Grain yield and related traits	4,302	1162	0.39-0.48	154
Oat	Heading date	635	13954	0.42-0.67	158

custom sequencing primers, and cost-ineffectiveness, and poor scalability of the current techniques⁹⁸. Nevertheless, if managing these factors, especially the lower cost and more scalability, becomes possible in near future, the haplotagging will be greatly used in the genetic studies. For instance, it will enable the haplotyping of the larger plant and animal populations, and allow the sequencing and systematic discovery of haplotypes in tens of thousands of samples, that too in both model and non-model plant species. It has been documented that both standard Illumina sequencing and haplotagging maintain full compatibility, and there is no extra cost in the haplotagging^{98,99}. The utility of haplotagging technique, for the identification of the haplotypes in the genome, has not yet been demonstrated in the plants, but recently, the haplotagging has been demonstrated in the two butterfly species⁸⁵. For example, Meier et al.⁸⁵ applied haplotagging approach to generate the haplotypes of megabase-size for the case of around six hundred butterflies' individuals belonging to the two species viz., *Heliconius erato* and *H. melpomene*, and these two species were identified to form hybrid zones that are overlapping across an elevational gradient in Ecuador. Besides, Meier et al.⁸⁵ also showed that haplotagging was able to detect the genetic loci regulating the distinct wing color patterns, namely, high- and low-land. In both the species the different haplotype alleles were detected at the same major loci; however, the chromosome rearrangements show no parallelism. To this end, this study demonstrated that technique of the "haplotagging" was successful to identify the distinct haplotype allele classes regulating the different phenotypes of the wing color patterns. Hence, these results suggested the enhanced power of the efficient haplotyping methods when combined with large-scale sequencing data from natural populations⁸⁵.

The above findings suggest the potential role of haplotagging in the identification of haplotype alleles regulating different phenotypes for a particular trait of interest. Hence, the haplotagging technique might be a promising strategy to identify the superior haplotype alleles in the diverse plant populations/germplasm for their ultimate use in the breeding for the development of improved crop varieties. This technique will be crucial to harness the true potential of the haplotype-based breeding for crop improvement.

Haplotype vs. individual markers: Comparative efficiency for crop breeding. Variations in the complex phenotypes are associated with the presence of SNPs, insertion-deletions and copy number variations in certain genomic loci^{100–102}. Currently, most of the plant breeders are using SNP markers to tag novel genetic variations underlying different phenotypes, and introgress these variations into the elite crop cultivars. However, the superiority of haplotype markers compared to individual SNP markers in addressing complex traits has been demonstrated through efficient gene identification and GS²⁶. For example, the use of haplotypes has been reported to considerably increase the prediction accuracy of the low-heritable quantitative traits as compared to the individual SNP markers^{103–107}. Besides, the use of haplotypes in gene mapping analyses has emerged as a more efficient

approach for the identification of genomic loci and candidate genes regulating traits of interest^{72,108,109}. The latest evidence suggests that the haplotype-based approach can improve not only the predictive abilities of GS models but also the precision with which genomic loci are detected in GWAS^{109–111}.

The higher efficiency of the haplotypes over individual SNP is due to some important reasons. For example, SNPs tiled on arrays are usually chosen for their moderate to high minor allele frequency (MAF). Therefore, most of the SNPs in the commercial chips are expected to be the old mutations, given that all new mutations remain at a low frequency in the beginning and a large part of them may disappear before reaching considerable frequency¹¹². Since the single-nucleotide-based genomic relationship matrix (G_{SNP}) is based on SNPs with relatively high MAF, this may imply that G_{SNP} traces old relationships from distant relatives and, therefore, may trace less accurately the changes due to recent selection as compared to the multi-locus haplotype-based relationship matrix, G_{HAP} ¹¹². Meuwissen et al.¹¹² suggested that building the relationship matrix using haplotypes instead of single SNPs may improve the accuracy of genomic predictions. Another potential limitation of G_{SNP} is that the SNPs are biallelic and, therefore, their polymorphism information content (PIC) value is not high. This restricts the ability to effectively capture LD between SNPs and multi-allelic QTLs. On the other hand, haplotype blocks are generally "multi-allelic" and may therefore better capture LD with multi-allelic QTLs compared to individual SNPs¹¹². It is also worth noting that longer haplotype blocks provide more information about possible recent mutations and close relationships than the shorter ones^{113,114}. Furthermore, haplotype effects could also factor in local epistatic effects among QTLs located within the haplotype blocks¹¹³. In addition, G_{HAP} can differentiate between identical by descent (IBD) and identical by state (IBS), while G_{SNP} cannot. This is because long shared haplotype blocks are likely to come from common ancestors. Therefore, long haplotype blocks can better capture information on IBD regions than individual SNPs in GS experiments¹¹⁵.

Applications of haplotypes in genetic analysis and breeding

Gene mapping. Recent studies elucidate the great potential of GWAS for the genetic dissection of important traits in major crop species. Researchers have mostly used SNP markers for the GWAS analysis¹¹⁶, because of the ability of the NGS-based genotyping systems to provide genome-wide marker data in cost- and time-efficient manner¹¹. As mentioned earlier, SNP markers are biallelic in nature having low informativeness and mutational rate¹¹⁷. Besides, the SNP arrays possess the inherent ascertainment biases, and thus in the GWAS analyses, the significant SNPs often do not represent the causal molecular variants^{5,8}. It can be explained by the fact that rare alleles often determine the extreme phenotypes²³. The existence of LD between true molecular variant and the non-causative markers causes stronger marker-trait linkage than that of causal variant itself^{25,118}.

Several researchers advocate for using haplotypes for conducting GWAS (Fig. 2). Recent GWA studies based on empirical and

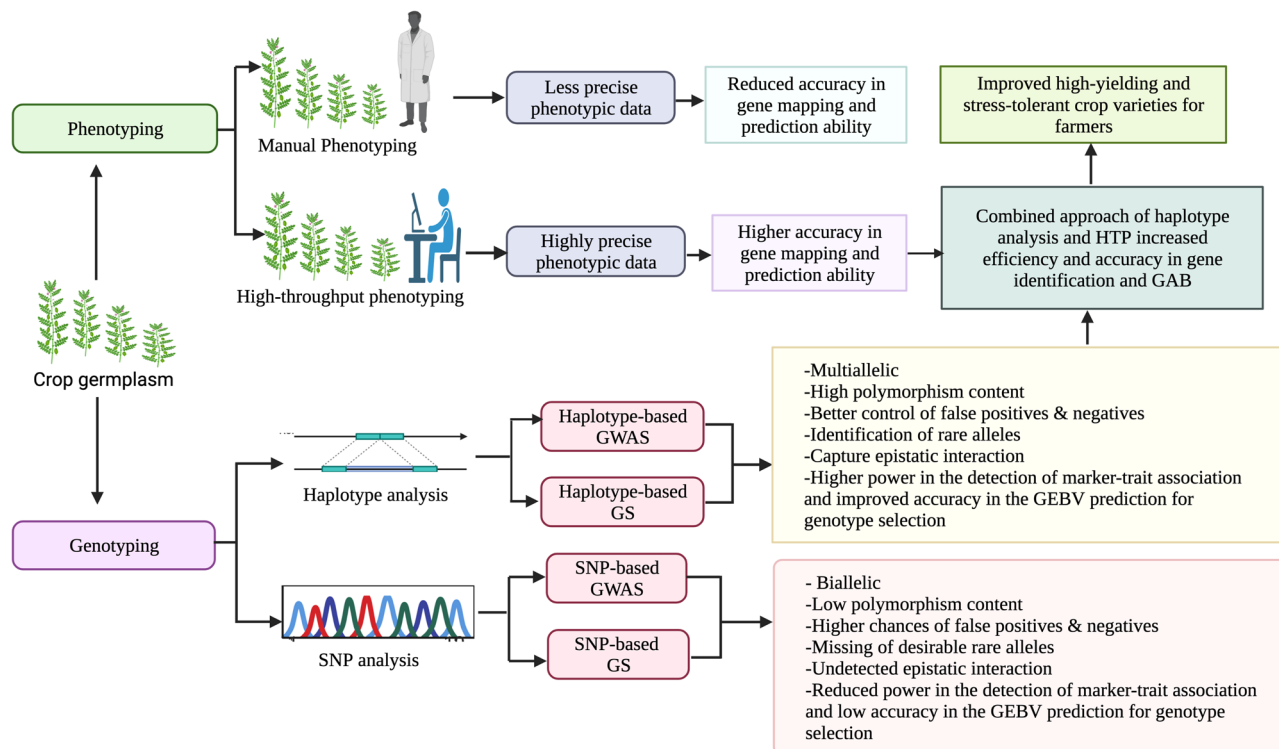


Fig. 2 Mining of SNPs and construction of haplotypes for detecting marker-trait associations (GWAS) and computing genomic estimated breeding values (GS). This diagram describes the comparative potential of the Haplotype-Based GWAS/Haplotype-Based GS in relation to SNP-Based GWAS/SNP-Based GS for the development of improved crop cultivars via genomics-assisted breeding (GAB). It showed that Haplotype-Based GWAS/Haplotype-Based GS in combination with the high-throughput phenotyping (HTP) has great potential to enhance the precision and accuracy in the gene identification and GAB. The image was created using BioRender (<https://biorender.com/>).

simulation data have revealed higher mapping accuracy and power of haplotype blocks over individual SNPs for the detection of QTLs/genes^{76,119–122}. A variety of reasons explain this superiority of haplotypes (Fig. 2). For example, Stephens et al.²⁷ demonstrated that the multi-allelic nature of haplotype blocks makes them more informative compared to SNP markers (biallelic in nature). The authors reported higher abundance of haplotype variants than SNPs, indicating recombination and recurrent mutation events within and among the genes in the haplotype. Moreover, the haplotype-based analysis is expected to control false positives and reveal the complex mechanism of causal haplotypes in a better way as compared to individual SNPs. For example, the repulsion states between two causal QTLs located close to each other²⁶. In particular, haplotype-based analysis can capture epistatic interactions between SNPs at a locus^{123,124}, provide more information to estimate whether two alleles are IBD¹²⁵, assess the biological role played by neighboring amino-acids on a protein structure¹²³, reduce the number of tests and hence the type I error rate¹²⁶, capture information from evolutionary history¹²⁷, and can provide more power than single marker system to analyze an allelic series existing at a particular locus^{128–131}. To this end, Hamblin and Jannink¹²⁹ reported that as compared to individual-based SNP markers, the haplotype approach increased the allelic effect and phenotypic variation explained (PVE) by 34% and 50%, respectively. N'Diaye et al.¹²⁰ observed that by combining multiple SNPs into haplotype blocks, the average PIC increased from 0.27 per SNP to 0.50 per haplotype in wheat. Over the last few years, haplotype-based GWAS analyses have identified important QTLs and candidate genes for various crop traits (Table 1). Greater power of haplotype-based mapping compared to SNP-based GWAS in the detection of genetic loci associated with the plant height and

biomass was evident in maize¹¹⁹. It is interesting to note that in comparison to single SNP-based mapping the haplotype-based mapping detected fewer significant associations and candidate genes for drought tolerance in maize; however, with higher PVE values¹³². Recently, applications of haplotype-based GWAS for various traits including yield, quality and stress tolerance in different plant species such as *Arabidopsis*¹³³, soybean¹³⁴, wheat¹²¹, barley^{131,135}, rice¹³⁶ and maize¹³⁷ have shown great promise for trait discovery and crop improvement.

However, the presence of non-informative SNPs in a given haplotype block (either small or long block) masks the effect of adjacent informative SNPs, which in turn leads to spurious associations, decreasing the effectiveness of the GWAS analysis¹³⁸. Hence, the haplotype-based GWAS and GS analyses uses the approaches such as sliding windows of fixed/variable length, haplotypes diversity among samples, LD between adjacent SNPs, and SNP number within haplotype to construct the haplotype blocks¹³⁹. All these approaches have one thing in common i.e., they all use the consecutive SNPs that possess high LD for the development of haplotypes. Therefore, under many circumstances, the haplotypes generated via these approaches' have been observed to show no difference in the information provided by the haplotype and single SNP, because the SNPs in high LD provide redundant information¹⁴⁰. To this end, recently a new haplotype-based GWAS approach called FH-GWAS has been introduced⁷⁶. This approach uses a different method to generate haplotypes i.e., only those SNPs are combined into functional haplotypes that possess true contribution to the haplotype effects via additive and/or epistatic effects. Thus, FH-GWAS is able to overcome the constraints of combining redundant SNPs (in high LD) into haplotypes and avoids the highly time-consuming process of selecting optimal combinations

of SNPs. It is therefore expected to be more powerful than SNP-based and other haplotype-based GWAS approaches.

FH-GWAS analysis: an efficient substitute for discovering superior haplotype alleles. Notwithstanding the superiority of GWAS based on haplotypes over SNPs, the use of haplotypes in the GWAS faces some challenges¹⁴¹. For instance, the contrasting effects of different haplotype allele classes will be diluted if the irrelevant markers are added to a possible causal genetic variant¹²³. Theoretically, in the case of a haplotype with m SNPs, the total number of different haplotype alleles will be equal to 2^m . This will increase the degree of freedom (this holds good for the estimation of population structure but not for GWAS, especially in the estimation of means and variance if the haplotypes are identified only once or twice), and that in turn will diminish the power of association analysis¹³¹. However, the 2^m formula for determining the number of haplotype alleles do not always work in practice because haplotype diversity is affected by a variety of factors including genetic structure and size of the population, mutation, recombination, marker ascertainment and demography¹⁴². For example, Scott et al.¹⁴³ by analyzing a panel of 16 wheat genotypes, representing the founders of MAGIC population, established that by using the SNPs of the promoter and genic regions, at most of the genes no greater than three haplotypes are identified, and most of the genes were biallelic. Besides, the most

critical factor affecting the haplotype-based GWAS analysis is the method(s) used for the construction of haplotypes, as discussed in the previous section. Only the consecutive SNPs in high LD are grouped into the haplotypes in all these methods. Sometimes the redundant information is provided by the SNPs that are in high LD, and as a result the use of these haplotypes does not provide more information than the individual SNPs¹⁴⁰. This explains the contradictions reported in recent studies regarding the efficiency of haplotype- and SNP-based GWAS approaches⁷⁶. As discussed above, the alternative approaches have been proposed for the identification of the haplotypes with non-consecutive SNPs that provide more information than the haplotypes with consecutive SNPs^{74,140,144}. Also, high computational burden associated with these approaches, further limits their use in the association studies⁷⁴.

To alleviate the limitations of the haplotype-based GWAS, an alternative efficient approach based on functional haplotype-based-GWAS (FH-GWAS) has been introduced to identify the superior haplotype alleles for the trait of interest⁷⁶ (Fig. 3). Given the significant role that the epistasis plays in the regulation of complex trait variations, FH-GWAS takes the associated epistatic effects of SNPs into consideration for trait discovery^{24,145,146}. Hence, in FH-GWAS, the SNPs possessing mild threshold for the main effects are first selected, followed by the identification of consecutive and/or non-consecutive combinations of SNPs

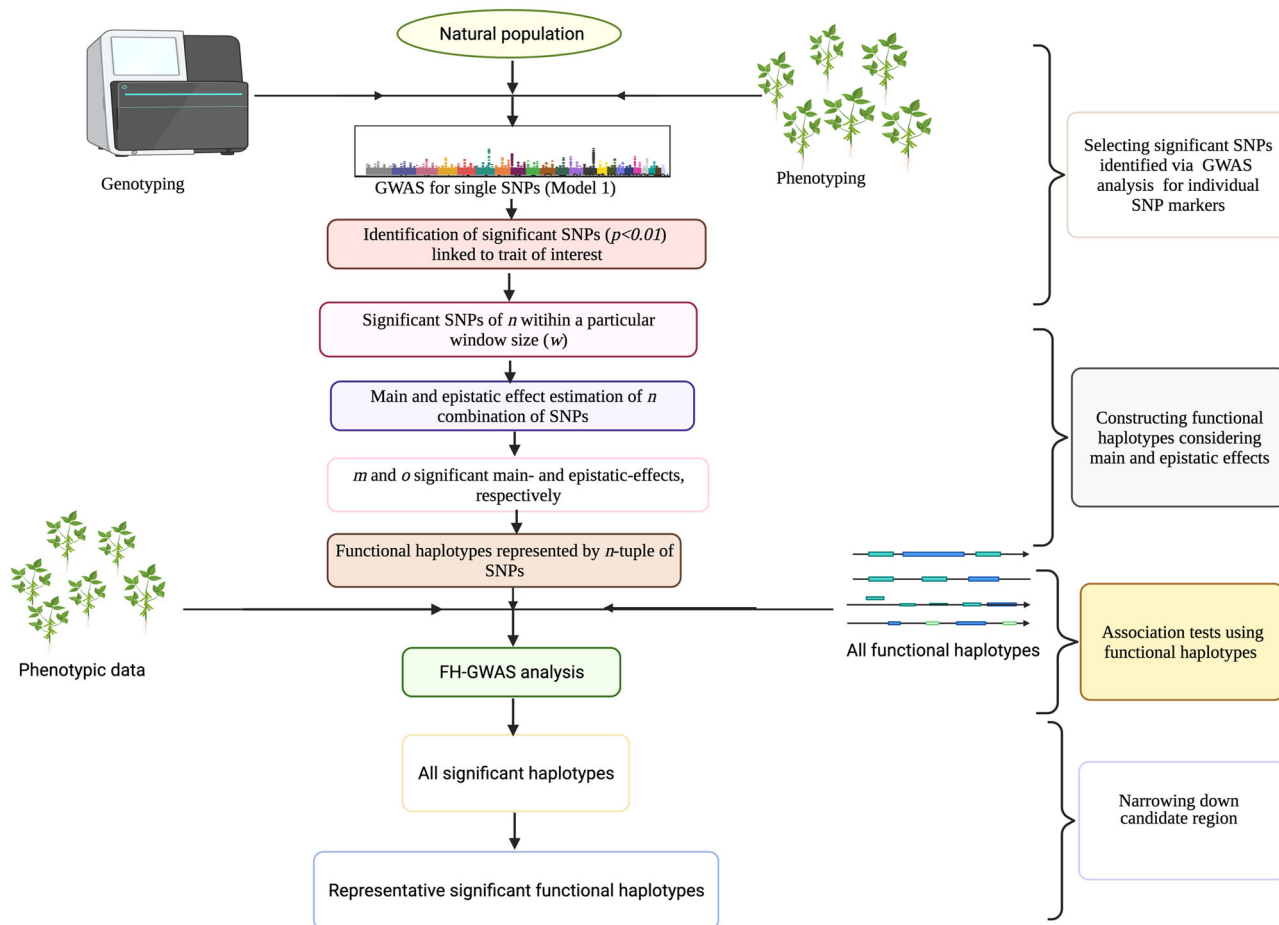


Fig. 3 Functional haplotype-GWAS (FH-GWAS) analysis for identification of the superior haplotypes for traits of interest. FH-GWAS approach first involves the individual SNP-marker based GWAS analysis (Model 1), that allows the identification of the candidate SNPs (SNP-trait association). This step is followed by the identification of the closely linked adjacent SNPs within a specific range in a chromosome region; and the SNPs within this specific region possessing additive and/or epistatic effects as well as have true contribution to the haplotype effects are combined into the functional haplotype. Lastly, the GWAS analysis was again performed by combining the functional haplotypes and phenotypic data, that ultimately leads to the identification of significant haplotypes associated with the trait of interest. The image was created using BioRender (<https://biorender.com/>).

(having significant epistatic effects) in a chromosomal region of defined size (Fig. 3). This approach combines only those SNPs into a functional haplotype that really contribute to the haplotype effects via additive and/or epistatic effects, thus preventing the redundant SNPs (with high LD) from combining into a haplotype. Besides, it prevents the laborious and time-consuming search for the detection of the optimal combinations of SNPs. In this regard, FH-GWAS is more powerful and efficient compared to haplotype-based and SNP-based approaches.

FH-GWAS outperformed SNP-based approach in a simulation study unless the SNPs of the haplotypes possess low MAF and the LD of haplotype SNPs is high⁷⁶. Analysis of flowering-time trait in a large population of *Arabidopsis thaliana* using FH-GWAS has revealed its great potential and efficiency in the association studies⁷⁶. Importantly, FH-GWAS detected all the genomic/candidate regions that were also identified via the SNP-based and haplotype-based GWAS approaches; however, it was only the FH-GWAS that could find a novel genomic region for flowering time on chromosome 4 of *A. thaliana*⁷⁶. In view of the evidences available from both simulation and empirical studies, FH-GWAS arguably holds a great promise for trait mapping in crop breeding. Further, this approach can be used for any crop species, particularly the homozygous ones, where sufficient coverage and suitable size of SNPs are available⁷⁶. However, if the FH-GWAS is to be used for the improvement of multiple traits, the construction of functional haplotypes for each individual trait must be done separately, as the tests of main and epistatic effects of markers are trait-dependent. Although FH-GWAS can improve the efficiency of the gene-trait association studies, this approach is computationally demanding in comparison to the other haplotype-based approaches⁷⁶.

Haplotype-based breeding (HBB). The development of stress-tolerant crop varieties with improved yield potential is one of the major challenges for breeders, especially in the face of global climate change^{3,124}. As discussed earlier, GS has emerged as an efficient approach for addressing complex polygenic traits, population improvement and developing improved varieties. The germplasm pool of the most crop species possesses complex genome structure; hence, the use of haplotypes in GS has been proposed as a powerful approach to improve the accuracy and efficiency in the prediction ability²⁶. This is because the comprehensive haplotype maps allow the identification and utilization of genomic regions linked to a particular trait at higher accuracy in populations with pronounced LD structures⁴.

Implementation of haplotypes in crop improvement is accomplished through two approaches, viz., retrospective and prospective⁸¹. During the long-term selection process, the plant breeders have selected the favorable haplotypes that lead to desirable phenotype(s) for the trait(s) of interest. Hence, by using the genome resequencing approach to sequence an elite gene pool, these favorable haplotypes can be identified in the elite crop germplasm²⁶. Furthermore, the molecular markers that define these favorable haplotypes can be developed and then all these haplotype-defining markers can be used to select the most desirable combination of haplotypes governing the specific phenotype. Besides, these haplotype-related markers can be used to separate favorable and unfavorable genetic variation by identifying lines with novel recombination in chromosomal blocks of interest. On the other hand, the haplotypes can also be used in the prospective manner, in which the large collection of ancestral and wild germplasm of particular crop species (not only the elite breeding pools) can be re-sequenced to identify haplotypes with a broader range of genetic variation⁸¹. In this approach, the genome-wide haplotypes are used to identify the novel haplotypes present in the wide range of natural germplasm.

Hence, the main objective of this approach is to identify the new, desirable and superior haplotypes. In summary, based on information/utility of various haplotypes, it is possible for assembling desirable haplotype combinations to develop optimal parents in breeding programmes. Deployment of haplotypes in breeding as mentioned above has been referred as haplotype-based breeding (HBB)^{6,20}.

Haplotype-assisted genomic selection. The prediction accuracies of GS models for yield and stress-related traits have outperformed the classical selection models, implying that GS is particularly suitable for the improvement of high-yielding and stress-tolerant crop cultivars^{3,147}. For example, Zhang et al.¹⁴⁸ demonstrated higher prediction accuracy of GS (0.75–0.87) as compared to MAS (0.62–0.75) for important agronomic traits in soybean. Similarly, GS was found superior to phenotypic selection for improving multiple agronomic traits related to yield and stress tolerance in different crop species¹⁴⁷. Besides, GS can reduce the time required to complete a selection cycle in crop plants, which can lead to increased production of the commercially important crops^{7,149}. Because of their high PIC value, fitting haplotypes with statistically significant associations to phenotypes as fixed effects in GS models could further improve prediction accuracies^{150,151}. The haplotype-assisted GS depicts the complex relationships between genotypic information and phenotypes more accurately than individual SNPs. Hence, this approach could ultimately help further increasing selection gain per unit of time. The use of haplotypes may improve the accuracy of genomic prediction because haplotypes can better capture LD and genomic similarity in different lines and may capture local high-order allelic interactions¹⁰⁹. Additionally, prediction accuracy could be improved by portraying population structure in the calibration set. A recent GS study that compared the prediction ability computed from haplotypes and SNPs in a set of 383 advanced lines and cultivars of wheat established the superiority of haplotype-based predictions over SNP-based predictions for all studied traits i.e., yield, test weight and protein content¹⁵². As compared to the individual SNPs, the combined use of haplotypes of 15 adjacent markers and training population optimization significantly improved the predictive ability for yield and protein content by 14.3% (four percentage points) and 16.8% (seven percentage points), respectively. Similar results were reported by other researchers in different crops such as maize¹⁵¹, *Brassica napus*¹⁵², and sorghum⁸⁰. Recent examples on the use of haplotype markers for genomic selection/prediction analysis in different crop species are presented in Table 2. Taken together, these studies underscore better performance of haplotypes in comparison to individual markers in improving prediction accuracies of GS for complex traits. Hence, the use of haplotypes in GS will definitely increase the prediction ability and greatly assist in harnessing the true potential of GAB in crop improvement.

Conclusion

GAB approaches aim to accelerate the pace of genetic gain and contribute to the global food and nutrition security. Several GAB approaches such as MABC, MARS and more recently GS have been successfully utilized for developing superior varieties. However, in the context of large-scale genome resequencing projects of germplasm accessions and breeding lines, it is possible to define new haplotypes. The availability of long-read sequencing technologies is also accelerating the discovery of haplotypes that are helpful to improve genome assembly. From applications perspective, these haplotypes can be used for a variety of purposes. Instead of using SNPs, haplotype-based GWAS analysis identifies causal polymorphism in a precise manner. Similarly,

evidence demonstrating higher genomic prediction efficiency, based on haplotypes as compared to SNPs, encourages researchers to increasingly embrace haplotypes-assisted genomic prediction in crop improvement programmes. Furthermore, advances in high-throughput phenotyping would enhance discovery and subsequent applications of superior haplotypes in crop breeding. We believe that haplotype-based research and their applications will be routine to develop improved cultivars for future food security.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Received: 21 May 2021; Accepted: 9 October 2021;
Published online: 04 November 2021

References

- Bhat, J. A. et al. Role of silicon in mitigation of heavy metal stresses in cplants. *Plants* **8**, 71 (2019).
- Ganie, S. A. & Reddy, A. S. N. Stress-induced changes in alternative splicing landscape in rice: Functional significance of splice isoforms in stress tolerance. *Biology* **10**, 309 (2021).
- Bhat, J. A. et al. Harnessing high-throughput phenotyping and genotyping for enhanced drought tolerance in crop plants. *J. Biotechnol.* **324**, 248–260 (2020).
- Varshney, R. K., Graner, A. & Sorrells, M. E. Genomics-assisted breeding for crop improvement. *Trends Plant Sci.* **10**, 621–630 (2005). **This comprehensive review describes that genomics research provides new genetic tools for crop improvement, which can in turn lead to the gradual evolution of the genomics-assisted breeding.**
- Bhat, J. A. et al. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Front. Genet.* **7**, 221 (2016).
- Varshney, R. K. et al. Designing future crops: genomics-assisted breeding comes of age. *Trends Plant Sci.* **26**, 631–649 (2021).
- Varshney, R. K., Terauchi, R. & McCouch, S. R. Harvesting the promising fruits of genomics: Applying genome sequencing technologies to crop breeding. *PLoS Biol.* **12**, e1001883 (2014).
- Zargar, S. M. et al. Recent advances in molecular marker techniques: insight into QTL mapping, GWAS and genomic selection in plants. *J. Crop Sci. Biotechnol.* **18**, 293–308 (2015).
- Przewieslik-Allen, A. M. et al. Developing a high-throughput SNP-based marker system to facilitate the introgression of traits from *Aegilops* species into bread wheat (*Triticum aestivum*). *Front. Plant Sci.* **9**, 1993 (2019).
- Huang, X. & Han, B. Natural variations and genome-wide association studies in crop plants. *Ann. Rev. Plant Biol.* **65**, 531–551 (2014). **The article illustrates how the advances in high-throughput sequencing technology will facilitate efficient use of crop diversity in the crop designs via genomics-assisted breeding.**
- Rasheed, A. et al. Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Mol. Plant* **10**, 1047–1064 (2017). **This review thoroughly discusses the scientific bottlenecks and overcoming strategies in the existing SNP-genotyping platforms, as well as their applications in crop improvement.**
- Ganal, M. W. et al. Large SNP arrays for genotyping in crop plants. *J. Biosci.* **37**, 821–828 (2012).
- Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R. & Crossa, J. Breeding schemes for the implementation of genomic selection in wheat (*Triticum spp.*). *Plant Sci.* **242**, 23–36 (2016).
- Yu, Z. et al. Identification of QTN and candidate gene for seed-flooding tolerance in soybean [*Glycine max* (L.) Merr.] using genome-wide association study (GWAS). *Genes* **10**, 957 (2019).
- Robertson, C. D., Hjortshøj, R. L. & Janss, L. L. Genomic selection in cereal breeding. *Agronomy* **9**, 95 (2019).
- Voss-Fels, K. & Snowden, R. J. Understanding and utilizing crop genome diversity via high-resolution genotyping. *Plant Biotechnol. J.* **14**, 1086–1094 (2016). **This review illustrates the role of next-generation sequencing and high-throughput SNP genotyping platforms in harnessing the untapped potential of crop diversity in crop improvement.**
- Collard, B. C., Jahufer, M. Z. Z., Brouwer, J. B. & Pang, E. C. K. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica* **142**, 169–196 (2005).
- Brachi, B., Morris, G. P. & Borevitz, J. O. Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.* **12**, 1–8 (2011).
- Zhu, C., Gore, M., Buckler, E. S. & Yu, J. Status and prospects of association mapping in plants. *Plant Genome* **1**, 5–20 (2008).
- Varshney, R. K. et al. 5Gs for crop genetic improvement. *Curr. Opin. Plant Biol.* **56**, 190–196 (2020).
- Crossa, J. et al. Genomic selection in plant breeding: Methods, models, and perspectives. *Trend. Plant Sci.* **22**, 961–975 (2017). **This review describes the principles and basis of genomic selection, and also gives a detailed account of statistical complexities/challenges associated with the estimation of GEBVs via different genomic prediction models.**
- Annicchiarico, P. et al. GBS-based genomic selection for pea grain yield under severe terminal drought. *Plant Genome* **10**, <https://doi.org/10.3835/plantgenome2016.07.0072> (2017).
- Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
- Mackay, T. F., Stone, E. A. & Ayroles, J. F. The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* **10**, 565–577 (2009).
- Korte, A. & Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Meth.* **9**, 1–9 (2013).
- Qian, L. et al. Exploring and harnessing haplotype diversity to improve yield stability in crops. *Front. Plant Sci.* **8**, 1534 (2017).
- Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
- Lu, J. et al. Mitochondrial haplotypes may modulate the phenotypic manifestation of the deafness-associated 12S rRNA 1555A>G mutation. *Mitochondrion* **10**, 69–81 (2010).
- Ganie, S. A., Wani, S. H., Henry, R. & Hensel, G. Improving rice salt tolerance by Precision Breeding in a New era. *Curr. Opin. Plant Biol.* **60**, 101996 (2021).
- Ceballos, H., Kawuki, R. S., Gracen, V. E., Yencho, G. C. & Hershey, C. H. Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. *Theor. Appl. Genet.* **128**, 1647–1667 (2015).
- Bradshaw, J. E. Plant breeding: past, present and future. *Euphytica* **213**, 60 (2017).
- Lenaerts, B., Collard, B. C. Y. & Demont, M. Review: improving global food security through accelerated plant breeding. *Plant Sci.* **287**, 110207 (2019).
- Banziger, M. & Diallo, A. O. Progress in developing drought and N stress tolerant maize cultivars for eastern and southern Africa in Integrated approaches to higher maize productivity in the new millennium. In *Proc. 7th Eastern and Southern Africa Regional Maize Conference, CIMMYT/KARI, Nairobi, Kenya* (eds Friesen, D. K. & Palmer, A. F. E.) 189–194 (CIMMYT (International Maize and Wheat Improvement Center) and KARI (Kenya Agricultural Research Institute, 2004).
- Qian, L., Qian, W. & Snowden, R. J. Haplotype hitchhiking promotes trait co-selection in *Brassica napus*. *Plant Biotechnol. J.* **14**, 1578–1588 (2016).
- Mühlaisen, J., Maurer, H. P., Stiewe, G., Bury, P. & Reif, J. C. Hybrid breeding in barley. *Crop Sci.* **53**, 819 (2013).
- Dong, H., Li, W., Tang, W. & Zhang, D. Development of hybrid Bt cotton in China—a successful integration of transgenic technology and conventional techniques. *Curr. Sci.* **86**, 778–782 (2004).
- Atlin, G. N., Cairns, J. E. & Das, B. Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change. *Glob. Food Sec.* **12**, 31–37 (2017). **This review emphasizes on the dire need of strengthened breeding system with short breeding cycles, high selection intensity, higher accuracy for cultivar development supported by genomic and phenomic technologies, and free international elite varietal exchange.**
- Labroo, M. R., Studer, A. J. & Rutkoski, J. E. Heterosis and hybrid crop breeding: a multidisciplinary review. *Front. Genet.* **12**, 643761 (2021).
- Khush, G. S. Rice breeding: past, present and future. *J. Genet.* **66**, 195–216 (1987).
- Ashraf, M. Inducing drought tolerance in plants: recent advances. *Biotechnol. Adv.* **28**, 169–183 (2010).
- Glenn, K. C. et al. Bringing new plant varieties to market: Plant breeding and selection practices advance beneficial characteristics while minimizing unintended changes. *Crop Sci.* **57**, 2906 (2017).
- Bradshaw, J. E. Review and analysis of limitations in ways to improve conventional potato breeding. *Potato Res.* **60**, 171–193 (2017).
- Saxena, R. K. et al. Genomics for greater efficiency in pigeonpea hybrid breeding. *Front. Plant Sci.* **6**, 793 (2015).
- Qaim, M. Role of new plant breeding technologies for food security and sustainable agricultural development. *Appl. Econom. Pers. Policy* **42**, 129–150 (2020).
- Evenson, R. E. Assessing the impact of the green revolution, 1960 to 2000. *Science* **300**, 758–762 (2003).

46. Sinha, P. et al. Superior haplotypes for haplotype-based breeding for drought tolerance in pigeonpea (*Cajanus cajan* L.). *Plant Biotechnol. J.* **18**, 2482–2490 (2020).
47. Varshney, R. K. et al. Fast-track introgression of root traits and other drought tolerance traits in JG 11, an elite and leading variety of chickpea. *Plant Genome* **6**, <https://doi.org/10.3835/plantgenome2013.07.0022> (2013).
48. Bharadwaj, C. et al. Introgression of “QTL-hotspot” region enhances drought tolerance and grain yield in three elite chickpea cultivars. *Plant Genome* **14**, e20076 (2021).
49. Henry, A., Gowda, V. R., Torres, R. O., McNally, K. L. & Serraj, R. Variation in root system architecture and drought response in rice (*Oryza sativa*): phenotyping of the OryzaSNP panel in rainfed lowland fields. *Field Crop Res.* **120**, 205–214 (2011).
50. Kumar, A. et al. Breeding high-yielding drought-tolerant rice: genetic variations and conventional and molecular approaches. *J. Exp. Bot.* **65**, 6265–6278 (2014).
51. Ahmed, H. U. et al. Genetic, physiological, and gene expression analyses reveal that multiple QTL enhance yield of rice mega-variety IR64 under drought. *PLoS ONE* **8**, e62795 (2013).
52. Henry, A. et al. Physiological mechanisms contributing to the QTL-combination effects on improved performance of IR64 rice NILs under drought. *J. Exp. Bot.* **66**, 1787–1799 (2015).
53. Hasan, M. M. et al. Marker-assisted backcrossing: a useful method for rice improvement. *Biotechnol. Biotechnol. Equip.* **29**, 237–254 (2015).
54. Cobb, J. N., Biswas, P. S. & Platten, J. D. Back to the future: revisiting MAS as a tool for modern plant breeding. *Theor. Appl. Genet.* **132**, 647–667 (2019). **This review discusses the potential of MAS in the modern crop breeding, and evaluates the processes needed for addressing the associated challenges.**
55. Dormatey, R. et al. Gene pyramiding for sustainable crop improvement against biotic and abiotic stresses. *Agronomy* **10**, 1255 (2020).
56. Gokidi, Y., Bhanu, A. N. & Singh, M. N. Marker assisted recurrent selection: an overview. *Adv. Life Sci.* **5**, 6493–6499 (2016).
57. Khan, A., Sovero, V. & Gemenet, D. Genome-assisted breeding for drought resistance. *Curr. Genomics* **17**, 330–342 (2016). 2016.
58. Ali, M. et al. Modeling and simulation of recurrent phenotypic and genomic selections in plant breeding under the presence of epistasis. *Crop J.* **8**, 866–877 (2020).
59. Borrell, A. K. et al. Drought adaptation of stay-green sorghum is associated with canopy development, leaf anatomy, root growth, and water uptake. *J. Exp. Bot.* **65**, 6251–6263 (2014).
60. Reddy, N. R. R., Ragimasalawada, M., Sabbavarapu, M. M., Nadoor, S. & Patil, J. V. Detection and validation of stay-green QTL in post-rainy sorghum involving widely adapted cultivar, M35-1 and a popular stay-green genotype B35. *BMC Genomics* **15**, 1–16 (2014).
61. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
62. Varshney, R. K. et al. Toward the sequence-based breeding in legumes in the post-genome sequencing era. *Theor. Appl. Genet.* **132**, 797–816 (2019).
63. Li, Y. et al. Investigating drought tolerance in chickpea using genome-wide association mapping and genomic selection based on whole-genome resequencing data. *Front. Plant Sci.* **9**, 190 (2018).
64. Beyene, Y. et al. Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Sci.* **55**, 154–163 (2015).
65. Juliana, P. et al. Retrospective quantitative genetic analysis and genomic prediction of global wheat yields. *Front. Plant Sci.* **11**, 580136 (2020).
66. Xu, Y. et al. Genomic selection of agronomic traits in hybrid rice using an NCII population. *Rice* **11**, 32 (2018).
67. Cui, Z. et al. Assessment of the potential for genomic selection to improve husk traits in maize. *G3* **10**, 3741–3749 (2020).
68. Stewart-Brown, B. B., Song, Q., Vaughn, J. N. & Li, Z. Genomic selection for yield and seed composition traits within an applied soybean breeding program. *G3* **9**, 2253–2265 (2019).
69. Roorkiwal, M. et al. Genomic-enabled prediction models using multi-environment trials to estimate the effect of genotype × environment interaction on prediction accuracy in chickpea. *Sci. Rep.* **8**, 11701 (2018).
70. Pandey, M. K. et al. Genome-based trait prediction in multi-environment breeding trials in groundnut. *Theor. Appl. Genet.* **133**, 3101–3117 (2020).
71. Stram, D. O. Multi-SNP haplotype analysis methods for association analysis. In *Statistical Human Genetics. Methods Mol. Biol.* (ed. Elston, R.) vol 1666, 485–504 (Humana Press, New York, NY, 2017).
72. Garg, S. Computational methods for chromosome-scale haplotype reconstruction. *Genome Biol.* **22**, 1–24 (2021). 2021.
73. Ammar, R., Paton, T. A., Torti, D., Shlien, A. & Bader, G. D. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Research* **4**, 17 (2015).
74. Maldonado, C., Mora, F., Scapim, C. A. & Coan, M. Genome-wide haplotype-based association analysis of key traits of plant lodging and architecture of maize identifies major determinants for leaf angle: Hap LA4. *PLoS ONE* **14**, e0212925 (2019).
75. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
76. Liu, F., Schmidt, R. H., Reif, J. C. & Jiang, Y. Selecting closely-linked SNPs based on local epistatic effects for haplotype construction improves power of association mapping. *G3* **9**, 4115–4126 (2019).
77. Huang, B. E., Amos, C. I. & Lin, D. Y. Detecting haplotype effects in genome wide association studies. *Genet. Epidemiol.* **31**, 803–812 (2007).
78. Gupta, P. K., Rustgi, S. & Kulwal, P. L. Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol. Biol.* **57**, 461–485 (2005).
79. Dixon, L. E., Pasquariello, M. & Boden, S. A. TEOSINTE BRANCHED1 regulates height and stem internode length in bread wheat. *J. Exp. Bot.* **71**, 4742–4750 (2020).
80. Jensen, S. M., Svendsgaard, J. & Ritz, C. Estimation of the harvest index and the relative water content—Two examples of composite variables in agronomy. *Eur. J. Agron.* **112**, 125962 (2020).
81. Bevan, M. W. et al. Genomic innovation for crop improvement. *Nature* **543**, 346–354 (2017). **This review illustrates the potential of modern genome sequencing platforms in the identification of wide spectrum of genes and genetic variations in crop plants, and explains how phenomics together with genomics has paved the way for the novel crop breeding systems.**
82. Abbai, R. et al. Haplotype analysis of key genes governing grain yield and quality traits across 3K RG panel reveals scope for the development of tailor-made rice with enhanced genetic gains. *Plant Biotechnol. J.* **17**, 1612–1622 (2019). **This study identifies the superior haplotypes for the traits related to grain yield and quality in rice using 3K genome panel and highlights the importance of haplotype-based breeding for developing next-generation tailor-made rice with superior haplotype combinations of target genes.**
83. Kalisz, S. & Kramer, E. M. Variation and constraint in plant evolution and development. *Heredity* **100**, 171–177 (2008).
84. Sella, G. & Barton, N. H. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.* **20**, 461–493 (2019).
85. Meier, J. I. et al. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *bioRxiv* <https://doi.org/10.1073/pnas.2015005118> (2020).
86. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nat. Rev. Genet.* **12**, 215–223 (2011).
87. Garud, N. R. & Rosenberg, N. A. Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps. *Theor. Popul. Biol.* **102**, 94–101 (2015).
88. Maestri, S. et al. A long-read sequencing approach for direct haplotype phasing in clinical settings. *Int. J. Mol. Sci.* **21**, 9177 (2020).
89. Delaneau, O. et al. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 1–10 (2019).
90. Amarasinghe, S. L. et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 1–16 (2020).
91. Al Bkhetan, Z., Zobel, J., Kowalczyk, A., Verspoor, K. & Goudey, B. 2019. Exploring effective approaches for haplotype block phasing. *BMC Bioinform.* **20**, 1–14 (2019).
92. Laver, T. W. et al. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Sci. Rep.* **6**, 1–6 (2016).
93. Zhang, S. et al. Long-read sequencing and haplotype linkage analysis enabled preimplantation genetic testing for patients carrying pathogenic inversions. *J. Med. Genet.* **56**, 741–749 (2019).
94. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).
95. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
96. Sun, H. et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *bioRxiv* <https://doi.org/10.1101/2021.05.15.444292> (2021).
97. Amini, S. et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).
98. Wang, O. et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.* **29**, 798–808 (2019).
99. Zhang, F. et al. Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat. Biotechnol.* **35**, 852–857 (2017).

100. Ganie, S. A., Molla, K. A., Henry, R. J., Bhat, K. V. & Mondal, T. K. Advances in understanding salt tolerance in rice. *Theor. Appl. Genet.* **132**, 851–870 (2019).
101. Khanzada, H. et al. Differentially evolved drought stress indices determine the genetic variation of Brassica napus at seedling traits by genome-wide association mapping. *J. Adv. Res.* **24**, 447–461 (2020).
102. Zhang, X. et al. Genetic variation in ZmTIP1 contributes to root hair elongation and drought tolerance in maize. *Plant Biotechnol. J.* **18**, 1271–1283 (2020).
103. Calus, M. P. et al. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genet. Sel. Evol.* **41**, 1–10 (2009).
104. Cuyabano, B. C., Su, G. & Lund, M. S. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics* **15**, 1–11 (2014). 2014.
105. Ballesta, P., Maldonado, C., Pérez-Rodríguez, P. & Mora, F. SNP and haplotype-based genomic selection of quantitative traits in Eucalyptus globulus. *Plants* **8**, 331 (2019).
106. Won, S. et al. Genomic prediction accuracy using haplotypes defined by size and hierarchical clustering based on linkage disequilibrium. *Front. Genet.* **11**, 134 (2020).
107. Matias, F. I., Galli, G., Correia Granato, I. S. & Fritsche-Neto, R. Genomic prediction of autogamous and allogamous plants by SNPs and haplotypes. *Crop Sci.* **57**, 2951–2958 (2017).
108. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
109. Hamazaki, K. & Iwata, H. RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. *PLoS Comput. Biol.* **16**, e1007663 (2020).
110. Vinholes, P., Rosado, R., Roberts, P., Borém, A. & Schuster, I. Single nucleotide polymorphism-based haplotypes associated with charcoal rot resistance in Brazilian soybean germplasm. *Agron. J.* **111**, 182–192 (2019).
111. Nyine, M. et al. Association genetics of bunch weight and its component traits in East African highland banana (Musa spp. AAA group). *Theor. Appl. Genet.* **132**, 3295–3308 (2019).
112. Meuwissen, T. H., Odegard, J., Andersen-Ranberg, I. & Grindflek, E. On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genet. Sel. Evol.* **46**, 1–8 (2014).
113. Hickey, J. M. et al. Sequencing millions of animals for genomic selection 2.0. *J. Anim. Breed. Genet.* **130**, 331–332 (2013).
114. Sargolzaei, M., Chesnais, J. P. & Schenkel, F. S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, 1–12 (2014).
115. Broman, K. W. & Weber, J. L. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am. J. Hum. Genet.* **65**, 1493–1500 (1999).
116. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).
117. Würschum, T., Maurer, H. P., Dreyer, F. & Reif, J. C. Effect of inter- and intragenic epistasis on the heritability of oil content in rapeseed (Brassica napus L.). *Theor. Appl. Genet.* **126**, 435–441 (2013).
118. Platt, A., Vilhjálmsson, B. J. & Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**, 1045–1052 (2010).
119. Lu, X. et al. Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. *Nat. Genet.* **44**, 890–894 (2012).
120. N'Diaye, A. et al. Single marker and haplotype-based association analysis of semolina and pasta colour in elite durum wheat breeding lines using a high-density consensus map. *PLoS ONE* **12**, e0170941 (2017).
121. Basile, S. M. L. et al. Haplotype block analysis of an Argentinean hexaploid wheat collection and GWAS for yield components and adaptation. *BMC Plant Biol.* **19**, 1–16 (2019).
122. Srivastava, R. K. et al. Genome-wide association studies and genomic selection in Pearl Millet: Advances and prospects. *Front. Genet.* **10**, 1389 (2020).
123. Clark, A. G. The role of haplotypes in candidate gene studies. *Genet. Epidemiol.* **27**, 321–333 (2004).
124. Bardel, C., Danjean, V., Hugot, J. P., Darlu, P. & Génin, E. On the use of haplotype phylogeny to detect disease susceptibility loci. *BMC Genet.* **6**, 1–13 (2005).
125. Meuwissen, T. H. E. & Goddard, M. E. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**, 421–430 (2000).
126. Zhao, K. et al. An Arabidopsis example of association mapping in structured samples. *PLoS Genet.* **3**, e4 (2007).
127. Templeton, A. R., Boerwinkle, E. & Sing, C. F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics* **117**, 343–351 (1987).
128. Akey, J., Jin, L. & Xiong, M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Euro. J. Hum. Genet.* **9**, 291–300 (2001).
129. Hamblin, M. T. & Jannink, J. L. Factors affecting the power of haplotype markers in association studies. *Plant Genome* **4**, <https://doi.org/10.3835/plantgenome2011.03.0008> (2011).
130. Morris, R. W. & Kaplan, N. L. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet. Epidemiol.* **23**, 221–233 (2002).
131. Gawenda, I., Thorwarth, P., Günther, T., Ordon, F. & Schmid, K. J. Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. *Plant Breed.* **134**, 28–39 (2015).
132. Yuan, X. & Biswas, S. Bivariate logistic Bayesian LASSO for detecting rare haplotype association with two correlated phenotypes. *Genet. Epidemiol.* **43**, 996–1017 (2019).
133. Lu, X. et al. Resequencing of cv CRI-12 family reveals haplotype block inheritance and recombination of agronomically important genes in artificial selection. *Plant Biotechnol. J.* **17**, 945–955 (2019).
134. Contreras-Soto, R. I. et al. A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. *PLoS ONE* **12**, e0171105 (2017).
135. Abed, A. & Belzile, F. Comparing single-SNP, multi-SNP, and haplotype-based approaches in association studies for major traits in Barley. *Plant Genome* **12**, 190036 (2019).
136. Wang, X. et al. Genome-wide and gene-based association mapping for rice eating and cooking characteristics and protein content. *Sci. Rep.* **7**, 1–10 (2017).
137. Yuan, Y. et al. Genome-wide association mapping and genomic prediction analyses reveal the genetic architecture of grain yield and flowering time under drought and heat stress conditions in maize. *Front. Plant Sci.* **9**, 1919 (2019).
138. Mathias, R. A. et al. A graphical assessment of p-values from sliding window haplotype tests of association to identify asthma susceptibility loci on chromosome 11q. *BMC Genet.* **7**, 1–11 (2006).
139. Srivastava, A. et al. Most frequent South Asian haplotypes of ACE2 share identity by descent with East Eurasian populations. *PLoS One* **15**, e0238255 (2020).
140. Laramie, J. M., Wilk, J. B., DeStefano, A. L. & Myers, R. H. HaploBuild: an algorithm to construct non-contiguous associated haplotypes in family based genetic studies. *Bioinformatics* **23**, 2190–2192 (2007).
141. Lorenz, A. J., Hamblin, M. T. & Jannink, J. L. Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PLoS ONE* **5**, e14079 (2010).
142. Stumpf, M. P. Haplotype diversity and SNP frequency dependence in the description of genetic variation. *Eur. J. Hum. Gene.* **12**, 469–477 (2004).
143. Scott, M. F. et al. Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding. *Genome Biol.* **22**, 1–30 (2021).
144. Knüppel, S. et al. Multi-locus stepwise regression: a haplotype-based algorithm for finding genetic associations applied to atopic dermatitis. *BMC Med Genet.* **13**, 8 (2012).
145. Carlborg, O. & Haley, C. S. Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.* **5**, 618–625 (2004).
146. Massawe, F., Mayes, S. & Cheng, A. Crop diversity: an unexploited treasure trove for food security. *Trend Plant Sci.* **21**, 365–368 (2016).
147. Matei, G. et al. Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. *Mol. Breed.* **38**, 1–13 (2018).
148. Zhang, J., Song, Q., Cregan, P. B. & Jiang, G. L. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (Glycine max). *Theor. Appl. Genet.* **129**, 117–130 (2016).
149. Qin, J. et al. Genome wide association study and genomic selection of amino acid concentrations in soybean seeds. *Front. Plant Sci.* **10**, 1445 (2019).
150. Jiang, Y., Schmidt, R. H. & Reif, J. C. Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *Genetics* **188**, 1687–1699 (2018). **This study demonstrates the higher prediction accuracies of haplotype-based model (HGBLUP) in comparison to marker-based models for the traits in a mouse panel, which suggests a great potential of haplotype-based breeding in genomic prediction.**
151. Sallam, A. H., Conley, E., Prakapenka, D., Da, Y. & Anderson, J. A. Improving prediction accuracy using multi-allelic haplotype prediction and training population optimization in wheat. *G3* **10**, 2265–2273 (2020).
152. Jan, H. U. et al. Genome-wide haplotype analysis improves trait predictions in Brassica napus hybrids. *Plant Sci.* **283**, 157–164 (2019).
153. Bruce, R. W. et al. Haplotype diversity underlying quantitative traits in Canadian soybean breeding germplasm. *Theor. Appl. Genet.* **133**, 1967–1976 (2020).
154. Sehgal, D. et al. Haplotype-based, genome-wide association study reveals stable genomic regions for grain yield in CIMMYT spring bread wheat. *Front. Genet.* **11**, 589490 (2020).

155. Ogawa, D. et al. Haplotype analysis from unmanned aerial vehicle imagery of rice MAGIC population for the trait dissection of biomass and plant architecture. *J. Exp. Bot.* **72**, 2371–2382 (2021).
156. Maldonado, C., Mora, F., Bertagna, F. A. B., Kuki, M. C. & Scapim, C. A. SNP- and haplotype-based GWAS of flowering-related traits in maize with network-assisted gene prioritization. *Agronomy* **9**, 725 (2019).
157. Mayer, M. et al. Discovery of beneficial haplotypes for complex traits in maize landraces. *Nat. Commun.* **11**, 4954 (2020). *The study identifies the haplotype-trait associations in ~1000 doubled haploid maize lines for early development traits and demonstrates that haplotype-based strategy has great potential for improving quantitative traits from genetic resources.*
158. Bekele, W. A., Wight, C. P., Chao, S., Howarth, C. J. & Tinker, N. A. Haplotype-based genotyping-by-sequencing in oat genome research. *Plant Biotechnol. J.* **16**, 1452–1463 (2018).
159. Ma, Y. et al. Potential of marker selection to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.). *Mol. Breed.* **36**, 113 (2016).
160. Jensen, S. E. et al. A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. *Plant Genome* **13**, e20009 (2020). *The Sorghum bicolor Practical Haplotype Graph (PHG) pangenome database, developed in this study demonstrated its utility in research and breeding.*

Acknowledgements

R.K.V. thanks Bill and Melinda Gates Foundation, USA (Grants ID# OPP1130244, OPP114827), Department of Cooperation and Farmers Welfare of the Ministry of Agriculture and Farmers Welfare, and JC Bose National Fellowship of Science & Engineering Research Board of Department of Science & Technology, Government of India for financial support. Authors are thankful to Rutwik Barmukh from ICRISAT for his help.

Author contributions

R.K.V. and S.A.G. conceptualized the idea and planned manuscript content. J.A.B. and D.Y. developed the manuscript draft. A.B. contributed special sections. A.B. and R.K.V. edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-021-02782-y>.

Correspondence and requests for materials should be addressed to Showkat Ahmad Ganie or Rajeev K. Varshney.

Peer review information *Communications Biology* thanks Hon-Ming Lam and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editors: Leena Tripathi and Caitlin Karniski.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021