Features and applications of haplotypes in crop breeding

Javaid Akhter Bhat, Deyue Yu, Abhishek Bohra, Showkat Ahmad Ganie, & Rajeev K. Varshney

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. 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 varieties, 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. This approach has delivered several high-yielding, stress-tolerant and better nutrition varieties. 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. 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 cost, and are now commercially available for both model and non-model crop species. These high-throughput platforms make hundreds of millions of DNA polymorphisms accessible for use in genetic and genomics research, and their application in crop breeding has considerably increased the gene mapping resolution and prediction accuracy in genomic selection (GS). 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) interactions.

1 National Center for Soybean Improvement, State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China. 2 Crop Improvement Division, ICAR-Indian Institute of Pulses Research (ICAR-IIPR), Kanpur, India. 3 Department of Biotechnology, Visva-Bharati, Santiniketan 731235 WB, India. 4 Center of Excellence in Genomics & Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad 502324, India. 5 State 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 approach.21

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 bi-allelic 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; rather 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 analyses.26 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 breeding.29 The science of plant breeding has evolved from conventional to present day GAB.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 breeding.44 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.1,6,46 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,31,52 and the improvement of different yield and stress-related traits in several major crops species.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 throughout the whole genome, rather than a few markers as in the case of MAS.61 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 profile.61 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 pipeline.61 Unlike MAS, GS does not necessarily require a prior knowledge of significant marker-trait associations.62 However, inclusion of the significant set of markers, such as resulting from GWAS, into GS models has been found to improve prediction accuracies.63 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.1,71,72 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 stretch.73 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 them\textsuperscript{74}. As shown in Fig. 1, two or more polymorphic SNPs of the haploid sequences inherited together as a unit constitute a haplotype\textsuperscript{71}. 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)\textsuperscript{75,76} and (c) by grouping of SNPs through sliding-windows of fixed or variable length\textsuperscript{77}. Evidence suggests that the LD-based approaches are more efficient for defining the haplotypes in the genomic/chromosome regions\textsuperscript{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 fragment\textsuperscript{78}.

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 selection\textsuperscript{26}. 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\textsuperscript{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 environments\textsuperscript{80}.

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.\textsuperscript{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-phenome analysis, Abbai et al.\textsuperscript{82} identified useful haplotypes for future breeding in rice and Sinha et al.\textsuperscript{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\textsuperscript{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 mutation\textsuperscript{3}. The trait-associated DNA markers are then used as surrogates for the selection of the desirable phenotypes\textsuperscript{5}. 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 studies\textsuperscript{85}, which prevents the accurate determination of ancestry reconstruction, rearrangements of chromosomes, allele-specific expression, and detection of selective sweeps\textsuperscript{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 variant\textsuperscript{88}. 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\textsuperscript{86,89}. 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 constructed\textsuperscript{88}. In comparison to the second-generation sequencing methods, analysis of DNA molecules can be performed directly via long-read sequencing platforms\textsuperscript{90}. 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 | PVE (%) | Haplotype-trait associations |
|--------------|-------|-----------------|-------------------|---------|----------------------------|
| Soybean      | 100-seed weight; plant height; seed yield | 169 | 941 | >10.0, 9.6-13.83 | 134 |
| Wheat        | Grain yield; days to heading; plant height | 6461 | 519 | 2.0-20.4 | 135 |
| Barley       | Deoxynivalenol content in kernels; heading time; days to maturity; grain yield; plant height; specific weight; 1000-kernel weight | 106 | 141 | 100.0 | 156 |
| Rice         | Grain shape | 372 | 372 | – | 155 |
| Maize        | Agronomic and reproductive traits | 322 | 322 | 5.6-22.0 | 74 |
| Maize        | Total plant height; ear height; ear height/plant height | >1000 | 183 | >10.0% | 157 |
| Rapeseed     | Days to flowering; seed glucosinolate content | 950 | – | – | 152 |

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. 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 variants. The shared barcode is then used to link the individual short reads for constructing the original haplotype. However, currently the commercial utilization of haplotagging in the genetic studies is prevented by certain factors, which include the requirement of.
custom sequencing primers, and cost-ineffectiveness, and poor scalability of the current techniques\textsuperscript{98}. 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\textsuperscript{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 species\textsuperscript{85}. For example, Meier et al.\textsuperscript{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.\textsuperscript{85} also showed that haplotagging was able to detect the genetic loci regulating the distinct wing color patterns, namely, high- and lowland. 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 populations\textsuperscript{85}.

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.

### 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 analysis\textsuperscript{116}, because of the ability of the NGS-based genotyping systems to provide genome-wide marker data in cost- and time-efficient manner\textsuperscript{11}. As mentioned earlier, SNP markers are biallelic in nature having low informativeness and mutational rate\textsuperscript{5,8}. It can be explained by the fact that rare alleles often determine the extreme phenotypes\textsuperscript{5,8}. The existence of LD between true molecular variants and the non-causative markers causes stronger marker-trait linkage than that of causal variant itself\textsuperscript{25,118}.

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

| 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              | 1,400             | >0.40                  | 151       |
| Wheat        | Grain yield and related traits              | 4,302            | 1,162             | 0.39-0.48              | 154       |
| Oat          | Heading date                               | 635              | 13954             | 0.42-0.67              | 158       |
Simulation data have revealed higher mapping accuracy and power of haplotype blocks over individual SNPs for the detection of QTLs/gene\textsuperscript{76,119–122}. A variety of reasons explain this superiority of haplotypes (Fig. 2). For example, Stephens et al.\textsuperscript{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 other\textsuperscript{26}. In particular, haplotype-based analysis can capture epistatic interactions between SNPs at a locus\textsuperscript{123,124}, provide more information to estimate whether two alleles are IBD\textsuperscript{125}, assess the biological role played by neighboring amino-acids on a protein structure\textsuperscript{123}, reduce the number of tests and hence the type I error rate\textsuperscript{126}, capture information from evolutionary history\textsuperscript{127}, and can provide more power than single marker system to construct the haplotype blocks\textsuperscript{128–131}. To this end, Hamblin and Jannink\textsuperscript{129} 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.\textsuperscript{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 maize\textsuperscript{119}. 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 values\textsuperscript{132}. Recently, applications of haplotype-based GWAS for various traits including yield, quality and stress tolerance in different plant species such as Arabidopsis\textsuperscript{133}, soybean\textsuperscript{134}, wheat\textsuperscript{121}, barley\textsuperscript{131,135}, rice\textsuperscript{136} and maize\textsuperscript{137} 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 analysis\textsuperscript{138}. 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 blocks\textsuperscript{139}. 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 information\textsuperscript{140}. To this end, recently a new haplotype-based GWAS approach called FH-GWAS has been introduced\textsuperscript{76}. 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.

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/).
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 challenges. For instance, the contrasting effects of different haplotype allele classes will be diluted if the irrelevant markers are added to a possible causal genetic variant. 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 analysis. 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 demography. For example, Scott et al. 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 SNPs. This explains the contradictions reported in recent studies regarding the efficiency of haplotype- and SNP-based GWAS approaches. 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. Also, high computational burden associated with these approaches, further limits their use in the association studies.

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 interest. FH-GWAS takes the associated epistatic effects of SNPs into consideration for trait discovery. 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/).
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 ability\(^26\). 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 structures\(^4\).

Implementation of haplotypes in crop improvement is accomplished through two approaches, viz., retrospective and prospective\(^81\). 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 germplasm\(^26\). 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 variation\(^81\). 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 species\(^4,47\). 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 accuracy\(^150,151\). 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 interactions\(^109\). 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 content\(^152\). 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 maize\(^151\), *Brassica napus*\(^82\), and sorghum\(^80\). 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 for 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 haplotype-assisted genomic prediction in crop improvement programs. 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**

1. Bhat, J. A. et al. Role of silicon in mitigation of heavy metal stresses in crops. *Plants* **8**, 71 (2019).
2. Genie, S. A. & Reddy, A. S. N. Stress-induced changes in alternative splicing landscape in rice: Functional significance of splice isoforms in stress tolerance. *Biotechnology* **10**, 309 (2021).
3. Bhat, J. A. et al. Harnessing high-throughput phenotyping and genotyping for enhanced drought tolerance in crop plants. *Journal of Biotechnology* **324**, 248–260 (2020).
4. Varshney, R. K., Grane, 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.
5. Bhat, J. A. et al. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Front. Genet.* **7**, 232 (2016).
6. Varshney, R. K. et al. Designing future crops: genomics-assisted breeding comes of age. *Trends Plant Sci.* **26**, 631–649 (2021).
7. 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).
8. 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).
9. 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).
10. Huang, X. & Han, B. Natural variations and genome-wide association studies in crop plants. *Ann. Rev. Plant Biol.* **65**, 531–551 (2014). This 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.
11. Rsahed, A. et al. Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Mol. Plant* **10**, 1047–1064 (2017). This review thoroughly discusses how advancements in genotyping platforms can help in the implementation of advanced breeding schemes.
12. 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).
13. Bradow, J. Plant breeding: past, present and future. *Euphytica* **213**, 60 (2017).
14. Lenars, B., Collard, B. C. Y. & Demont, M. Review: improving global food security through accelerated plant breeding. *Plant Sci.* **287**, 11207 (2019).
15. Banziger, M. & Di alo, 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).
16. Qian, L. Qian, W. & Snowden, R. J. Haplotypic hitchhiking promotes trait co-segregation in Brassica napus. *Plant Biotechnol. J.* **14**, 1578–1588 (2016).
17. Mushahwar, I. J., Maurer, H. P., Stewé, G., Bury, P. & Reif, J. C. Hybrid breeding in barley. *Crop Sci.* **53**, 819 (2013).
18. 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.* **96**, 778–782 (2004).
19. 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. *Global Food Sec.* **12**, 31–37 (2017). This review emphasizes 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.
20. Labrou, M. R., Studer, A. J. & Ruthkoski, J. E. Heterosis and hybrid crop breeding: a multidisciplinary review. *Front. Genet.* **12**, 643761 (2021).
21. Khush, G. S. Rice breeding: past, present and future. *J. Genet.* **66**, 195–216 (1987).
22. Ashraf, M. Inducing drought tolerance in plants: recent advances. *Biotecnol. Adv.* **28**, 169–183 (2010).
23. 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).
24. Bradshaw, J. E. Review and analysis of limitations in ways to improve conventional potato breeding. *Potato Res.* **60**, 171–193 (2017).
25. Saxena, R. K. et al. Genotyping for greater efficiency in pigeonpea hybrid breeding. *Front. Plant Sci.* **6**, 793 (2015).
26. Quim, M. Role of new plant breeding technologies for food security and sustainable agricultural development. *Appl. Econom. Pers. Policy* **42**, 129–150 (2020).
27. Evenson, R. E. Assessing the impact of the green revolution, 1960 to 2000. *Science* **300**, 758–762 (2003).
59. Borrell, A. K. et al. Drought adaptation of stay-green sorghum is associated with canopy development, leaf anatomy, root growth, and water uptake. *Exp. Bot.* 120, 205–214 (2011).

60. Kumar, A. et al. Breeding high-yielding drought-tolerant rice: genetic variations and conventional and molecular approaches. *J. Exp. Bot.* 65, 2625–2627 (2014).

61. 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, e57725 (2013).

62. Varshney, R. K. et al. Toward the sequence-based breeding in legumes in the post-genome sequencing era. *Theor. Appl. Genet.* 133, 9–22 (2014).

63. 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 ofwide spectrum of genes and genetic variations in crop plants, and explains how phenomics together with modern genomics has paved the way for the novel crop breeding systems.

64. Abbas, 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.* 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.

65. Dormatey, R. et al. Gene pyramiding for sustainable crop improvement against biotic and abiotic stresses. *Agronomy* 10, 1235 (2020).

66. Cui, Z. et al. Assessment of the potential for genomic selection to improve quality traits across 3K RG panel reveals scope for the development of tailor-made rice with enhanced genetic gains. *Plant Biotechnol.* 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.

67. Kalisz, S. & Kramer, E. M. Variation and constraint in plant evolution and biodiversity. *Am. J. Hum. Genet.* 223, 65–76 (2011).

68. Wang, Y. et al. Exploring effective approaches for haplotype block phasing. *BMC Bioinform.* 20, 1–14 (2019).

69. Laver, T. W. et al. Pitfalls of haplotype phasing from ampiclon-based long-read sequencing. *Sci. Rep.* 6, 1–6 (2016).

70. 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).

71. Weissenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* 27, 757–767 (2017).

72. Werner, 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).

73. Sun, H. et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *bioRxiv* https://doi.org/10.1101/2021.05.14.444292 (2021).

74. Amini, S. et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* 46, 1343–1349 (2014).

75. 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).

76. Zhang, F. et al. Haplotype phasing of whole human genomes using head-based barcode partitioning in a single tube. *Nat. Biotechnol.* 35, 852–857 (2017).
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 IC 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 ICRI SAT 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.

© The Author(s) 2021