Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CHAP-GWAS: Leveraging Chromosomal Haplotypes to Improve Genome-Wide Association Studies

Zhenyu Jia

arthur.jia@ucr.edu

University of California, Riverside https://orcid.org/0000-0002-1428-904X

Shibo Wang

University of California, Riverside

Qiong Jia

University of California, Riverside

Yanru Cui

Hebei Agricultural University

Han Qu

University of California, Riverside https://orcid.org/0000-0002-6016-5152

Ruidong Li

Graduate Program in Genetics, Genomics, and Bioinformatics, University of California

Lei Yu

Genetics, Genomics and Bioinformatics Graduate Program, University of California

Xuesong Wang

University of California, Riverside

Chin-Sheng Teng

University of California, Riverside

Si Liu

University of California, Riverside

Chenwu Xu

Agricultural college, Yangzhou University https://orcid.org/0000-0002-1971-1086

Yuan-Ming Zhang

Huazhong Agricultural University https://orcid.org/0000-0003-2317-2190

Sarah Wang

University of California, Riverside

Shuhan Yin

University of California, Riverside

Loren Ho

University of California, Riverside

Meiyue Wang
University of California, Riverside

Weiming Chen
University of California, Riverside

Weide Zhong
Department of Urology, School of Medicine, The Second Affiliated Hospital of South China University of Technology (Guangzhou First People's Hospital) https://orcid.org/0000-0002-0430-2845

Jianguo Zhu
guizhou provincial people's hospital https://orcid.org/0000-0002-1171-0126

Danelle Seymour
University of California, Riverside

Shou-Wei Ding
University of California, Riverside

Shizhong Xu
University of California, Riverside

Wenxiu Ma
University of California, Riverside https://orcid.org/0000-0003-4097-1621

Yang Xu
Agricultural college, Yangzhou University https://orcid.org/0000-0002-6493-4916

Article

Keywords:

Posted Date: August 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10333409/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

CHAP-GWAS: Leveraging Chromosomal Haplotypes to Improve Genome-Wide Association Studies

Shibo Wang[1†], Qiong Jia[1,2†], Yanru Cui[1,3†], Han Qu[1], Ruidong Li[1,2], Lei Yu[1,2], Xuesong Wang[1,2], Chin-Sheng Teng[4], Si Liu[5,6], Chenwu Xu[7], Yuanming Zhang[8], Sarah Wang[1], Shuhan Kelly Yin[1], Loren L. Ho[1], Meiyue Wang[1], Weiming Chen[1,9], Weide Zhong[10,11,12], Jianguo Zhu[9,13], Danelle Seymour[1], Shou-Wei Ding[5], Shizhong Xu[1], Wenxiu Ma[4*], Yang Xu[1,7*], Zhenyu Jia[1,2*]

1 Department of Botany and Plant Sciences, University of California, Riverside 92521, USA
2 Graduate Program in Genetics, Genomics, and Bioinformatics, University of California, Riverside 92521, USA
3 College of Agronomy, Hebei Agricultural University, Baoding 071001, China
4 Department of Statistics, University of California, Riverside 92521, USA
5 Department of Microbiology & Plant Pathology, University of California, Riverside 92521, USA
6 National Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, Hubei Hongshan Laboratory, Wuhan 430070, China
7 College of Agronomy, Yangzhou University, Yangzhou 225009, China
8 Crop Information Center, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, China
9 Medical School of Guizhou University, Guiyang, Guizhou 550031, China
10 Department of Urology, Guangdong Key Laboratory of Clinical Molecular Medicine and Diagnostics, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou 510180, China
11 Guangdong Key Laboratory of Urology, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou Medical University, Guangzhou 510120, China
12 Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macau 999078, China
13 Department of Urology, Guizhou Provincial People's Hospital, Guiyang, Guizhou 550002, China

Abstract

Conventional genome-wide association studies (GWAS), which analyze individual single nucleotide polymorphisms (SNPs), have been instrumental in dissecting the genetic basis of quantitative traits. While short-range haplotype variants have recently demonstrated superior statistical power by capturing local linkage disequilibrium, they are often constrained by predefined, fixed-length blocks. Advances in long-read sequencing and computational phasing are making chromosome-scale haplotypes increasingly accessible, opening new opportunities for association studies. However, frameworks that effectively exploit this long-range genetic information remain limited. Here, we introduce CHAP-GWAS, a novel, dynamically adaptive algorithm designed to identify trait-associated haplotype variants from chromosome-scale phased data. Unlike existing methods, CHAP-GWAS utilizes a phenotype-driven approach that dynamically defines and extends haplotype blocks from "seed" loci, allowing for the precise capture of complex genetic effects. Using comprehensive simulations and real-world datasets from three diverse plant species (*Arabidopsis*, rice, and maize), we demonstrate that CHAP-GWAS successfully identifies quantitative trait loci overlooked by conventional single-SNP and static haplotype-based approaches. By leveraging chromosome-length haplotypes, our framework significantly enhances statistical power and addresses a critical component of missing heritability. CHAP-GWAS offers a robust,

scalable, and synergistic approach to genetic mapping, with the potential to advance precision agriculture and medicine through more comprehensive and accurate genetic discovery.

Main (Introduction)

Conventional genome-wide association studies (GWAS), which rely on testing individual single-nucleotide polymorphisms (SNPs), have facilitated the identification of genomic loci associated with complex traits, ranging from human diseases to key agronomic traits. However, the detected loci often explain only a fraction of the heritability, leaving a substantial portion of the phenotypic variance unexplained – a phenomenon termed "missing heritability"[1,2]. Rare or small-effect genetic variants, incomplete linkage disequilibrium (LD), or ungenotyped genomic regions[3,4] potentially contribute to missing heritability. Haplotype variants, representing blocks of consecutive SNPs, capture local LD patterns more effectively than single-SNP tests[5,6] and increase imputation accuracy for ungenotyped regions[7], thereby improving the resolution of genetic variation[8] and overall statistical power.

Several strategies have been developed for haplotype phasing and local haplotype construction in haplotype-based GWAS. Statistical phasing tools such as BEAGLE[9], SHAPEIT[10], and Eagle[11] have been widely used for haplotype phase inference from genotype or sequencing data. Based on the phased sequences, local haplotypes can subsequently be constructed using fixed windows[12], LD-defined blocks identified by tools such as PLINK[13] and Haploview[14], or predefined annotation-based regions such as genes and coding regions[15]. However, these haplotype construction strategies generally rely on genotype information, local LD patterns, or predefined genomic boundaries, without using haplotype–phenotype associations to guide trait-specific haplotype construction or selection. Furthermore, computational demands and limited scalability remain major bottlenecks, particularly for long-range haplotypes. Recent sequencing technologies have enabled the generation of long and ultra-long reads, which directly capture haplotype information over extended genomic regions[16-18]. However, achieving accurate, genome-wide haplotype inference remains challenging due to sequencing gaps and coverage biases, ultimately limiting their utility in downstream genetic analyses. More recent methods, such as RAINBOW[19], utilize SNP-set models to analyze unphased chromosomal genotypic data to infer chromosome-wide haplotype blocks, followed by conventional GWAS. Nonetheless, these approaches still rely on predefined blocks, and haplotypes remain indistinguishable when constitutional SNPs are all heterozygous. This highlights the ongoing need for more adaptive frameworks that can handle chromosomal-scale phased data.

Homozygous populations, such as recombinant inbred lines (RILs), provide readily available and relatively stable chromosome-scale haplotypes for direct association testing. Furthermore, cost-efficient methods now enable the inference of chromosome-length haplotypes even in highly heterozygous species, such as humans and citrus[20]. These chromosome-range haplotypes more precisely capture the underlying genetic structure and facilitate the detection of important trait associations. In this study, we developed the Chromosomal Haplotype-Integrated GWAS (CHAP-GWAS), the first chromosome-scale haplotype-based GWAS framework leveraging chromosome-resolved genomic data. Unlike previous haplotype-based approaches that often rely on predefined LD blocks, CHAP-GWAS dynamically identifies and extends haplotype variants governed by phenotypic associations. We validated CHAP-GWAS through comprehensive simulations and real-data analyses across *Arabidopsis*, rice, and maize, demonstrating that QTL detection power increases significantly compared to conventional SNP- and predefined short-range haplotype-based models. By incorporating chromosome-scale haplotypes, this method helps

recover a portion of the missing heritability and provides a more comprehensive understanding of the genetic architecture of complex traits.

Table 1 Summary of Datasets

Species	Arabidopsis			Rice	Maize
Sample Size	199 (Homozygous accessions)	496 (Homozygous accessions)	500 (Homozygous accessions)	565 (F1 hybrids)	1000 (F1 hybrids)
Genome size (Mb)	119Mb	119Mb	119Mb	373 Mb	2060
SNP number	216,095	1,823,394	1,833,408	1,619,588	2,373,301
SNP density (10kb)	18.1	153.0	153.9	43.4	11.5
Number of phenotypes	6	1	1	2	1
Effective SNP number	Varies	283,740	292,381	121,863	2,373,301
Initial p value threshold	0.05	0.05	0.05	0.05	0.05
Final p value threshold	2.86e-06*	3.52e-06 #	3.42e-06 #	4.10e-07*	2.11e-08*

* Final p value threshold is calculated by 0.05/Effective SNP number, while for # this threshold was calculated by 1/Effective SNP number.

Result

Haplotype-Based GWAS Leveraging Phased Chromosomal SNPs

Fig. 1 outlines the core algorithm of CHAP-GWAS for haplotype identification analysis. In Step 1 (Fig. 1b), we use a sliding window to scan each di-SNP variable, the simplest haplotype consisting of two adjacent SNPs. As in conventional GWAS, each di-SNP is analyzed like an individual SNP, using a relaxed significance criterion (p-value < 0.05; see Materials and Methods). All di-SNP variables passing this screening are retained, even those with weak associations. They are then used as haplotype seeds for further haplotype extension. In Step 2 (Fig. 1b), each di-SNP seed is evaluated in both directions and extended along the one with the stronger association signal. Haplotype extension stops when adding an adjacent SNP to either side reduces the association signal. In this study, haplotype variables refer to DNA variants consisting of three or more consecutive SNPs. Once the final haplotype variables have been determined, each is tested using a stringent p-value cutoff adjusted for multiple comparisons (Fig. 1a). This new haplotype-based association analysis is integrated with conventional single-SNP association analysis in CHAP-GWAS. In the following sections, we analyze datasets for *Arabidopsis* (homozygous accessions), rice (F1 hybrids), and maize (F1 hybrids) (Table 1).

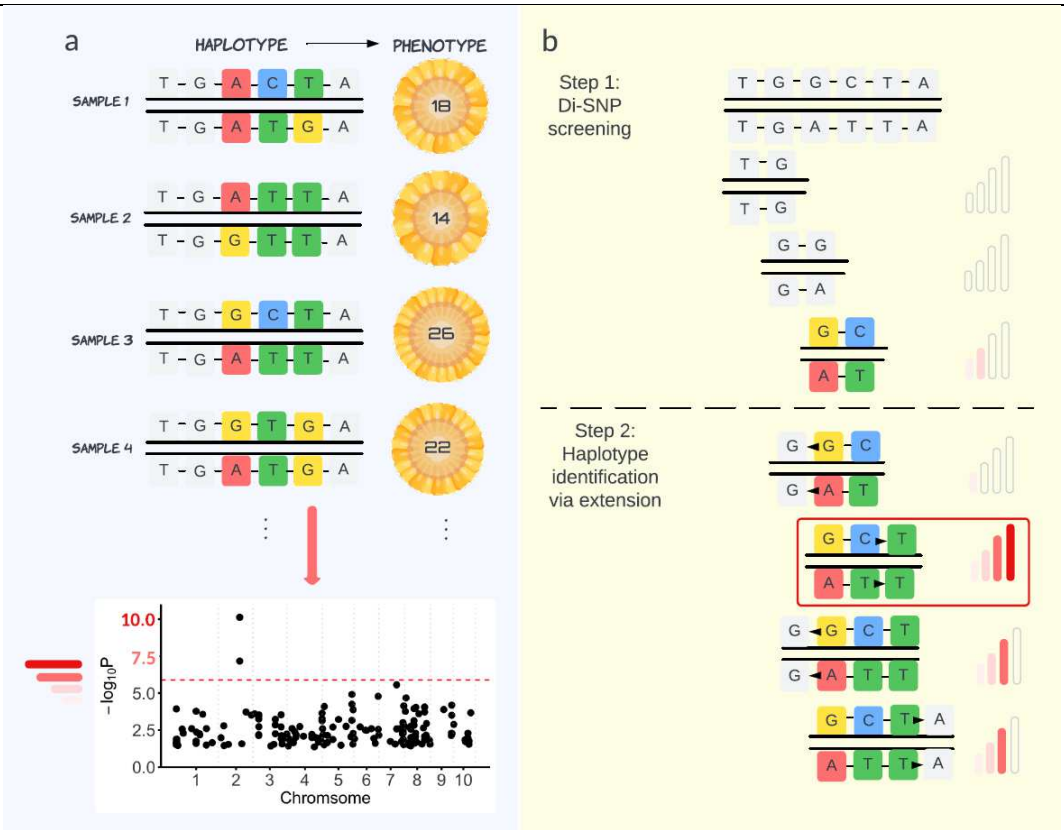


Fig. 1: Illustration of haplotype-based association analysis in maize. a, Upper panel: Example of an association between a detected haplotype and kernel number per row (KNR). Lower panel: Manhattan plot from haplotype-based GWAS showing chromosomal position (x axis) and association strength (y axis). The red dashed line indicates the selection threshold, and black dots above the threshold mark genomic loci of trait-associated haplotypes. **b,** Algorithm for identifying the trait-associated haplotypes. Upper panel (Step 1): Sliding-window screening of candidate di-SNP seeds. Lower panel (Step 2): Haplotype extension from di-SNP seeds until the association signal plateaus.

Analysis of *Arabidopsis* Agronomy Dataset

Due to the homozygosity of each accession, the SNPs on a chromosome can be used as phased chromosomal SNPs in CHAP-GWAS. We analyzed six phenotypes in *Arabidopsis* (Supplementary Table S1) using CHAP-GWAS to compare the performance of three types of DNA variants: single SNPs, di-SNPs, and haplotypes identified by the new 'haplotype growth' algorithm. Fig. 2 presents the results of analyzing 'days to flowering under long days for four weeks vernalization (4W)', an important ecological trait in *Arabidopsis*. Some loci are detected by a single type of DNA variant (single SNPs, di-SNPs, or haplotypes), while other loci may be identified by multiple types of DNA variants. For instance, Fig. 2a shows that one SNP and two di-SNPs associated with 4W are located within a 20 kb region to the right of the *SVP* gene (SHORT VEGETATIVE PHASE). *SVP* encodes a MADS-box transcription factor critical for regulating flowering time[21]. Fig. 2b illustrates one 4-SNP and one 5-SNP haplotype overlapping with the *LUH* gene (LEUNIG-HOMOLOG). This gene encodes a transcriptional co-repressor involved in regulating gene expression during various developmental processes, including floral development[22]. In Fig. 2c, one 3-SNP and one 4-SNP haplotype are located within a 20 kb region to the left of the *SAP* gene (Stress-Associated

Protein). *SAP* is expressed during flowering and seed development. It also plays roles in developmental regulation under stress conditions[23]. Fig. 2f–g further show that the identified haplotypes show stronger association signals for 4W than their individual SNP components. Fig. 2e shows that both di-SNPs are significant. The shared middle SNP exhibits a slightly stronger signal. This suggests that the central variant is the primary driver of the association. Therefore, conventional GWAS remains effective in detecting variants driven by single-nucleotide changes. Conversely, Fig. 2f–g illustrate that individual SNPs show low or nonsignificant associations with 4W. In contrast, haplotype associations are significant in both instances. Similar patterns were observed for the remaining five agronomic traits (Supplementary Figures S1 – S5). These results showed that (1) CHAP-GWAS outperformed RAINBOWR, and (2) CHAP-GWAS successfully identified novel associations overlooked by single-SNP GWAS, highlighting the complementary advantages of integrating both frameworks for a more exhaustive genetic analysis.

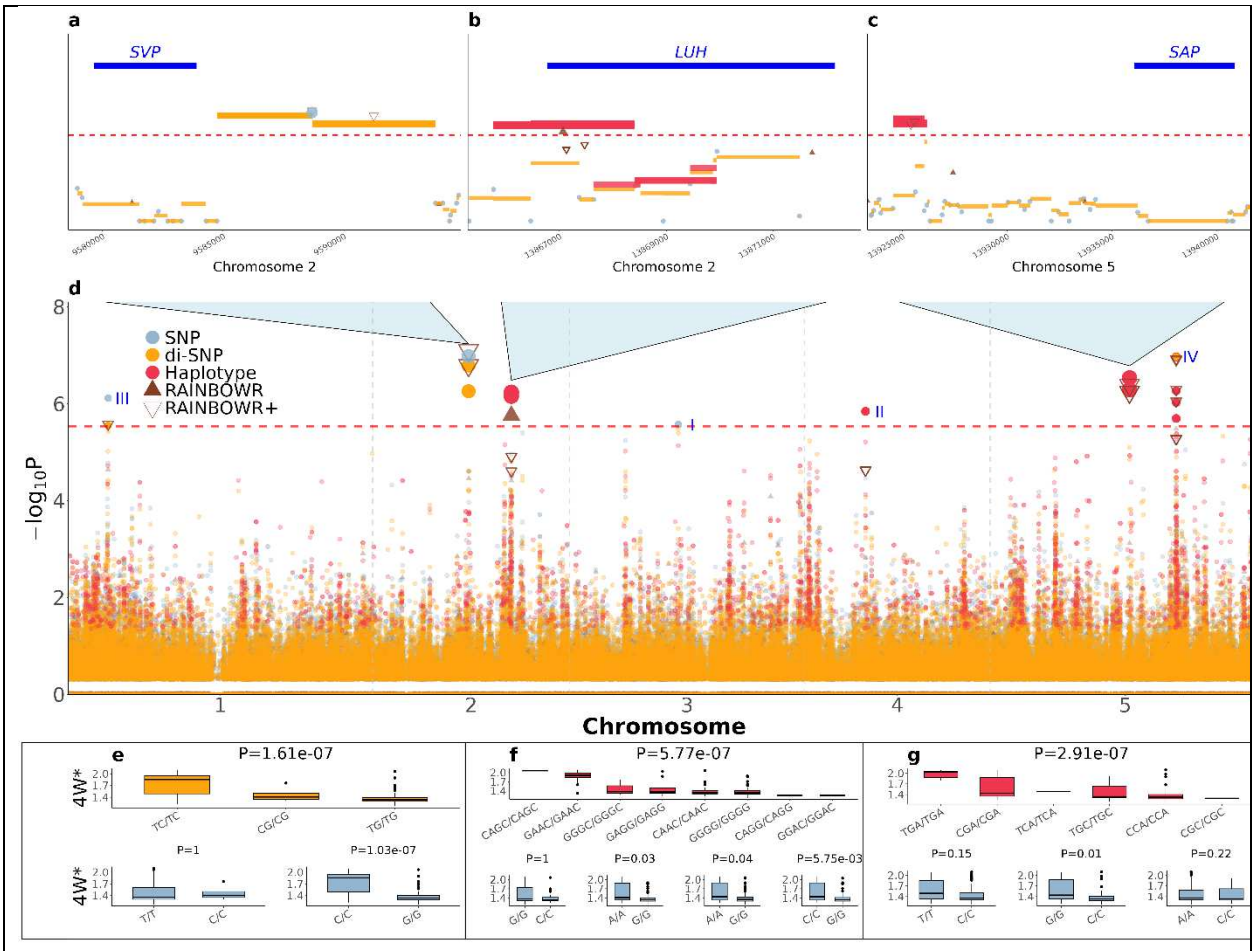


Fig. 2: Analysis of the log-transformed trait "days to flowering under long days" (4W*) in *Arabidopsis*. a–c, Zoom-in views of three genomic loci and their adjacent genes on chromosomes 2 and 5 identified through di-SNP and SNP analyses. **d,** Manhattan plot summarizing association results, distinguishing between single SNPs (blue), di-SNPs (orange), haplotypes (red), and RAINBOWR analyses using PLINK-pre-selected (solid brown triangles) and CHAP-GWAS-selected haplotypes (hollow brown triangles). Roman numerals I–IV categorize variant clusters as: I, exclusively SNP detection; II, exclusively haplotype detection; III, SNPs overlapping with di-SNPs or haplotypes; IV, haplotypes overlapping with di-SNPs. **e–g,** Comparison of the strongest haplotype signals with their constituent SNPs at the three genomic regions shown in **a–c**. The upper panels depict 4W* distributions across haplotypes, and the lower panels show 4W* distributions among genotypes of each SNP within these haplotypes.

Analysis of *Arabidopsis* Virus Datasets

We previously published GWAS results for two *Arabidopsis* resistance traits to endemic Cucumber Mosaic Virus (CMV): the accumulation levels of Q-CMV and CMV-Δ2b. These levels were quantified using RT-qPCR and ELISA, respectively[24]. Using conventional GWAS, we identified a limited number of associated loci under a relaxed significance threshold. We further characterized the two highest-ranking genomic loci: one on chromosome 4 for CMV-Δ2b and another on chromosome 5 for Q-CMV. Our experiments demonstrated the positive and negative regulatory roles in antiviral RNAi defense for two genes associated with these loci: Reduced Dormancy 5 (*RDO5*) and Antiviral RNAi Regulator 1 (*VIR1*). Fig. 3 demonstrates that haplotype variants identified these two loci with stronger association strength than single SNP variants (vertical rectangles in panels d and h). Additionally, haplotype variants uncovered several further loci that failed to meet the stringent significance thresholds required in single-SNP tests. These newly identified genomic regions, potentially associated with CMV resistance, include three regions on Chromosomes 1 and 4. These regions are tightly linked to genes *MYC2*[25], *NUA*[26], and *EDS5*[27], all of which are functionally relevant to virus resistance. **Supplementary Table S2** summarizes the number of loci detected for *Arabidopsis* resistance traits to CMV by different DNA variants. This case demonstrates that CHAP-GWAS enhances the existing GWAS results, yielding stronger evidence for known genes and revealing novel biological insights.

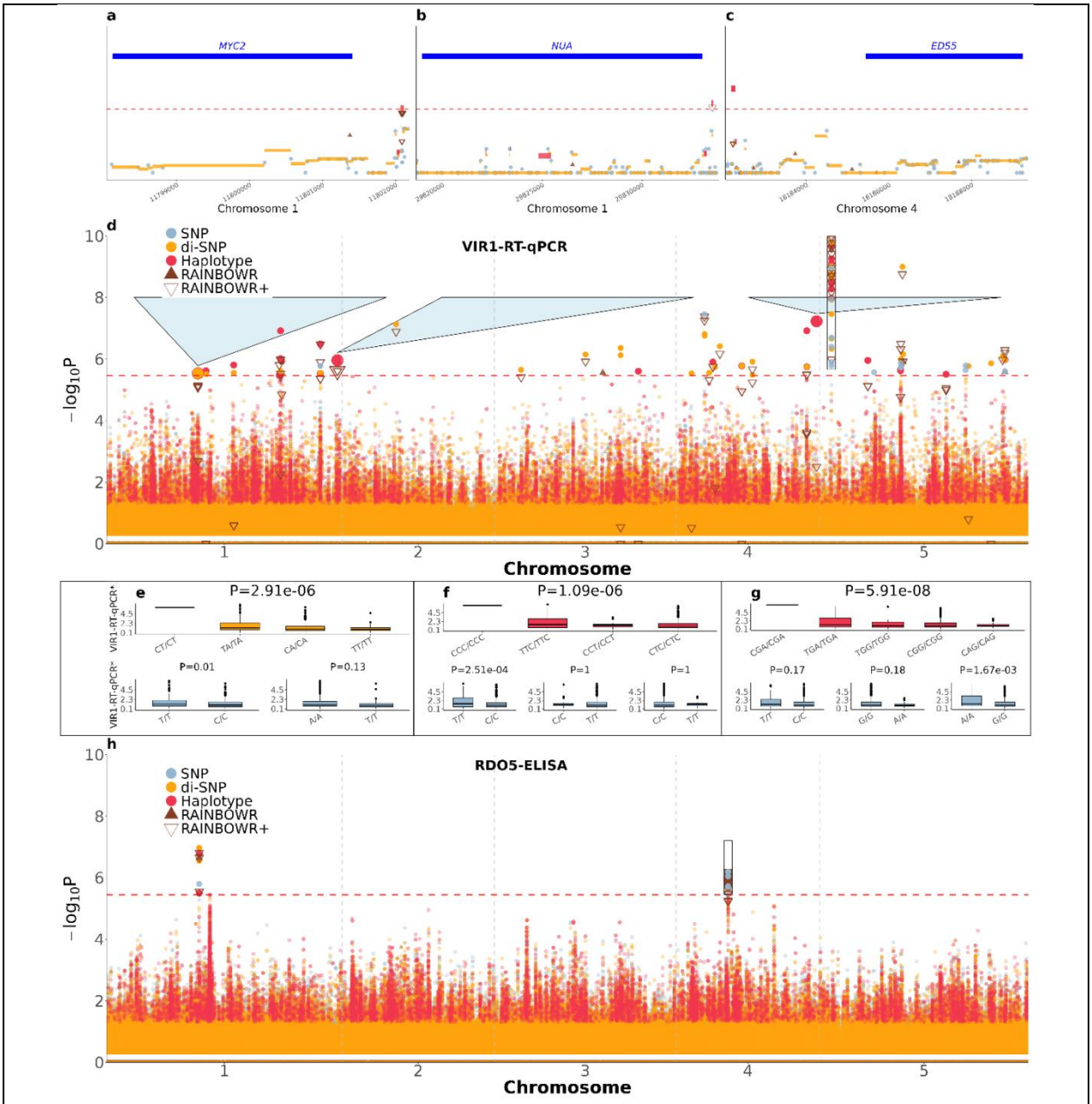


Fig. 3: Analysis of *Arabidopsis* resistance to endemic cucumber mosaic virus (CMV) accumulation. This figure presents the results for two *Arabidopsis* resistance traits to endemic CMV: Q-CMV accumulation (log-transformed as Q-CMV*) and CMV-Δ2b, quantified by RT–qPCR and ELISA, respectively. **a**–**c**, Zoom-in views of three loci and nearby genes on chromosomes 1 and 4 associated with Q-CMV, detected using di-SNP and haplotype analyses (**a**) or haplotype variants alone (**b**, **c**). **d**, Manhattan plot showing association results for Q-CMV based on single SNPs (blue), di-SNPs (orange), haplotypes (red), and RAINBOWR analyses using PLINK-pre-selected (solid brown triangles) and CHAP-GWAS-selected haplotypes (hollow brown triangles) (hollow brown triangles). A vertical rectangle highlights the genomic region on chromosome 5 that contains the Antiviral RNAi Regulator 1 (VIR1) gene. This gene was identified and validated in our previous report. **e–g,** Comparisons of Q-CMV association strengths of the strongest haplotypes and their constituent SNPs at the three loci are shown in **a**–**c**. Upper panels show Q-CMV distributions across haplotypes with overall p-values. Lower panels show Q-CMV distributions between genotypes at each SNP within the detected haplotype variants. **h**, Manhattan plot for CMV-Δ2b. A vertical rectangle highlights the genomic region on chromosome 4 containing the Reduced Dormancy 5 (RDO5) gene, which was also identified and validated in our previous report.

Analysis of Simulated *Arabidopsis* F1 Hybrids

To assess the increased statistical power of CHAP-GWAS using chromosomal haplotypes in hybrids, we simulated 200 F1 hybrids from 199 homozygous accessions. For this simulation, we used 'days to flowering under long days' (DFLD), another important flowering trait, for demonstration. As shown in Fig. 4a, each hybrid's background genome was formed by combining haploid genomes from two randomly sampled parental accessions, and the baseline phenotype (DFLD*) was their average parental value. Haplotypes, either 3 or 4 SNPs in length, were introduced at a selected genetically neutral locus *via* three distinct schemes: (1) The original DNA sequence was replaced by pre-designed haplotypes, ensuring that each constituent SNP is identically heterozygous across all simulated hybrids. (2) Natural DNA variation observed at the selected locus of the simulated hybrids was leveraged to define a haplotype variant. (3) The PLINK haplotype mining function[13] was applied to search for and define a haplotype variant based on the natural DNA variation observed in the simulated hybrids. Note that in scheme (1), conventional GWAS based on single SNP variants cannot detect any constituent SNPs within the simulated haplotype variant because there is no genetic variation across the hybrids at these single heterozygous SNP loci. In contrast, in schemes (2) and (3), there is genotype variation in single SNP loci within the simulated haplotype variant; therefore, these constituent SNPs may be detectable by conventional GWAS. In each scheme, the genetic effects associated with haplotypes with varying levels of broad-sense heritability were assigned and added to the baseline phenotype. Each scheme, haplotype length, and heritability combination formed a scenario, which was tested across different neutral genomic regions. To compare the average detection power (defined as the rate of successful detection of all component SNPs within a simulated trait-associated haplotype) across all replicates for each scenario, we analyzed three DNA variants: SNPs, di-SNPs, and haplotypes. For statistical power comparison, we also used a loose criterion for successful identification of the genomic region containing the simulated haplotype: detection of any constitutional SNPs of the haplotype. We also included analysis by RAINBOWR (a widely used method for haplotype-based GWAS), which used PLINK-pre-selected haplotypes as input. See details of the simulation procedures in **Materials and Methods**.

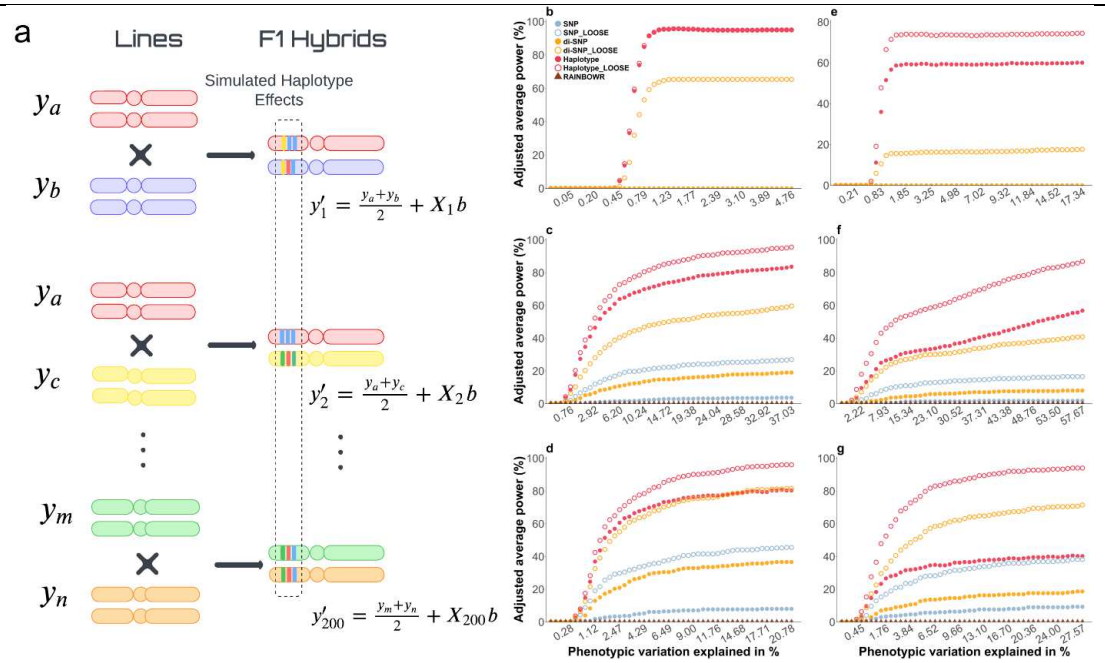


Fig. 4: Analysis of simulated Arabidopsis F_1 hybrids. a, Diagram illustrating how the genome and phenotype (DFLD*, representing the log-transformed "days to flowering under long days") were generated for each hybrid. The background genome for each hybrid, represented by two parental chromosomes for simplicity, was randomly sampled from 199 homozygous accessions. The baseline phenotype was calculated as the average of the phenotypes of the two sampled parental lines. Haplotype variables were introduced at selected loci through three distinct mechanisms. Scheme 1 (**b**, **e**), replacement of the original sequence at the selected locus with pseudo (pre-designed) haplotype sequences; Scheme 2 (**c**, **f**), definition of haplotypes directly from natural DNA variation at neutral loci; and Scheme 3 (**d**, **g**), use of PLINK to define haplotypes at neutral loci as blocks of consecutive SNPs. Three-SNP haplotypes were simulated in **b**–**d**, and four-SNP haplotypes in **e**–**g**.

As expected, statistical power increased with heritability for both 3-SNP and 4-SNP haplotypes (Fig. 4b–g). In scheme (1), due to the lack of genotype variation at any constituent SNPs of the simulated haplotype variants, none of them could be detected by conventional GWAS methods. For the same reason, RAINBOWR cannot detect these simulated haplotypes in these two simulations. Therefore, only the statistical power between di-SNPs and haplotypes was compared, demonstrating the overwhelming performance of haplotype variables over di-SNPs (Fig. 4b, e). In both schemes (2) and (3), we also observed that the use of haplotypes outperformed di-SNPs or single SNP variants in detecting the simulated trait-associated haplotype variants. Di-SNPs also surpassed single SNP variants, whereas RAINBOWR exhibited the lowest performance (Fig. 4c, d, f, g). These results were consistent for both 3-SNP and 4-SNP simulated haplotypes. While 3-SNP haplotypes had higher statistical power than 4-SNP at low heritability, statistical power converged at high heritability.

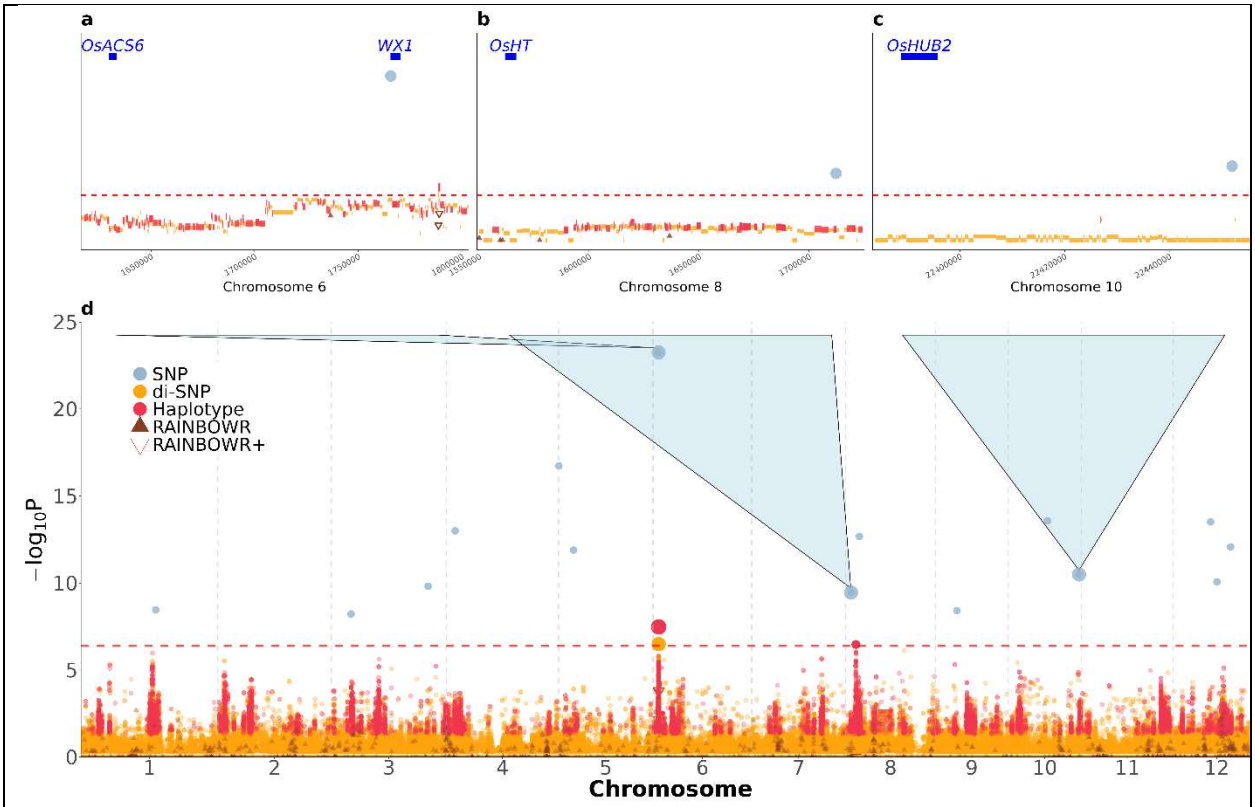


Fig. 5: Association analysis of amylose content (AC) in rice. a, Genomic region on chromosome 6 containing *OsACS6* and *WX1*, detected by two haplotype variants and one SNP. **b,** Genomic region on chromosome 8 containing *OsHT*, detected by a single SNP. **c,** Genomic region on chromosome 10 containing *OsHUB2*, also detected by a single SNP. **d,** Manhattan plot showing association results for AC, displaying single SNPs (blue), di-SNPs (orange), haplotypes (red), and RAINBOWR analyses using PLINK-pre-selected (solid brown triangles) and CHAP-GWAS-selected haplotypes (hollow brown triangles).

Analysis of Rice Quality Datasets

Using CHAP-GWAS, we compared SNPs, di-SNPs, and haplotypes for amylose content (AC) and protein content (PC) (Supplementary Table S3). Our analysis utilized data from 113 parental lines and 565 derived hybrids[28]. The original study employed a statistical adjustment in the conventional GWAS because it is difficult to quantify these traits. We therefore used their single-SNP analysis results as the baseline. This provided a direct benchmark for our di-SNP and haplotype variant analysis. Some genomic loci were detected by only one variant type, while others were detected by multiple types. For AC, two significant SNPs were associated with the known genes *OsHT*[29] on chromosome 8 and *OsHUB2*[30] on chromosome 10, respectively. Furthermore, a single region was detected by two haplotype variants and one SNP. This region is close to two known genes, *OsACS6*[31] and *WX1*[32] on chromosome 6. For PC, a single SNP detected a region on chromosome 3 associated with the known gene *GS3*[33]. In contrast, haplotype analysis identified two distinct regions in the vicinity of *RGG2*[34] and *GW2*[35] on chromosome 2. Overall, haplotype variants derived from our "haplotype growth" algorithm substantially improve the statistical power of conventional GWAS.

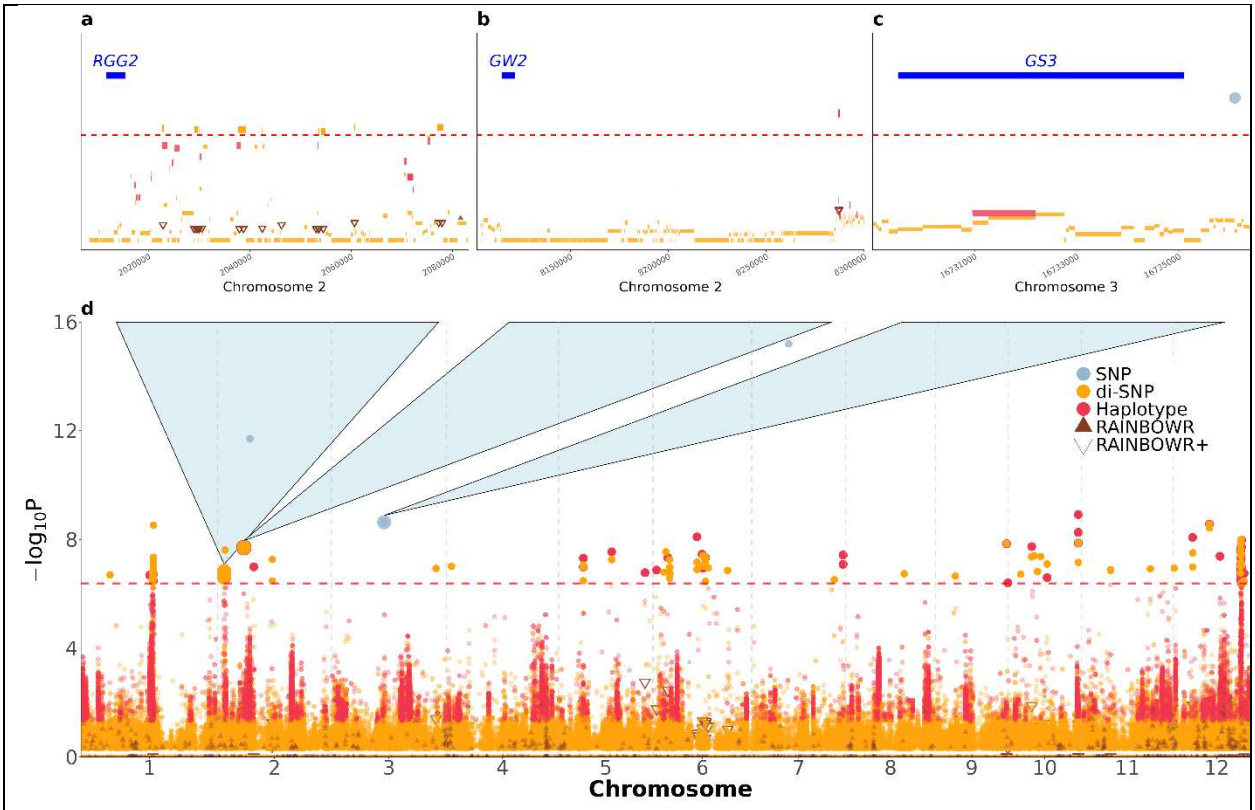


Fig. 6: Association analysis of protein content (PC) in rice. a, Genomic region on chromosome 2 containing *RGG2*, detected only by haplotype variants. **b,** Genomic region on chromosome 2 containing *GW2*, also detected only by haplotype variants. **c,** Genomic region on chromosome 3 containing *GS3*, detected only by a single SNP. **d,** Manhattan plot showing association results for PC, displaying single SNPs (blue), di-SNPs (orange), haplotypes (red), and RAINBOWR analyses using PLINK-pre-selected (solid brown triangles) and CHAP-GWAS-selected haplotypes (hollow brown triangles).

Analysis of Maize hybrid Datasets (semi-simulation)

Fig. 7 illustrates the results of association analysis for plant height (PH) in maize. We used a semi-simulation approach to evaluate the effects of sample size and SNP marker density on GWAS performance. Fig. 7c represents the baseline analysis based on all 1000 samples and over 2 million SNPs, yielding the highest statistical power and genome coverage. Reducing the sample size to 25% and 50% in panels (a) and (b), respectively, mainly lowered the height of association peaks and led to the loss of some signals. In contrast, reducing marker density to 50% and 25% in panels (d) and (e) significantly lowered genome coverage and mapping resolution. As a result, peaks became narrow, fragmented, or disappeared. Haplotype variants consistently outperformed single-SNP variants across all scenarios. These results showed that sufficient sample size and dense marker coverage are essential for maximal detection power and localization accuracy in CHAP-GWAS.

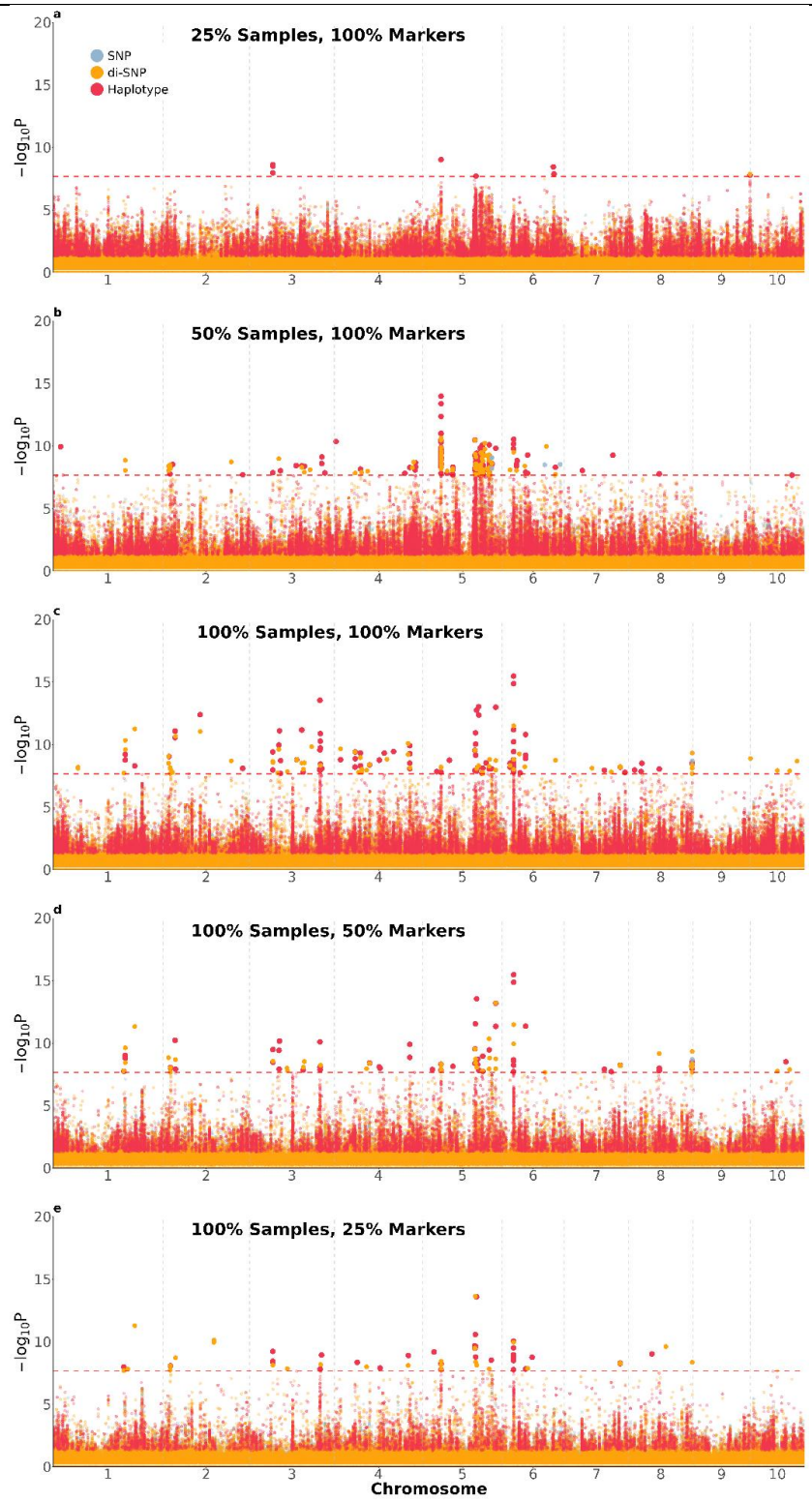


Fig. 7: Association analysis of plant height (PH) in maize using 1,000 randomly selected hybrids. a, b, Manhattan plots from analyses using all SNPs for 250 and 500 hybrids, respectively, randomly selected from the full set of 1,000 hybrids. **c,** Manhattan plot from the analysis of all SNPs for the complete set of 1,000 hybrids. **d, e,** Manhattan plots from analyses using 50% and 25% of randomly selected SNPs, respectively, for the 1,000 hybrids.

Comparative Analysis of SNP, Di-SNP, and Haplotype Associations

Unique SNP Findings

Single-SNP GWAS identified distinct loci missed by di-SNPs and haplotypes, particularly for *Arabidopsis* flowering traits. For example, SNPs located near *FLC*[36] and *AGL17*[37], both key regulators of flowering, were identified in trait DFLD. For virus resistance, a few unique SNPs were detected with no clear gene associations. In rice quality traits, significant SNPs were identified near *OsHT*[29] and *OsHUB2*[30] for AC, and near *GS3*[33] for PC. By contrast, no unique SNPs were detected in maize. These results reflect that the genetic architecture varies across species and traits.

Unique Haplotype Discoveries

Haplotype-based analysis uncovered several unique loci missed by SNP- or di-SNP-based approaches. In *Arabidopsis*, haplotypes near *GATA18*[38] and *LEUNIG*[39] (trait SD), *LUH*[22] (trait 4W), and *SAP*[23] (trait 4W), all critical for flower development, were detected. In virus resistance, haplotypes near *NUA*[26] and *EDS5*[27], both involved in plant immunity, were uniquely identified. For rice quality trait PC, di-SNPs – representing short haplotypes without extension – were found near *RGG2*[34]. These results highlight the enhanced sensitivity of haplotype analysis in capturing complex genetic associations often missed by single-SNP GWAS.

Overlapping SNP, di-SNP, and Haplotype Results

Substantial overlap was observed among SNPs, di-SNPs, and haplotypes. For *Arabidopsis* flowering traits, overlaps occurred near *SVP* and *JMJ27*. *SVP* represses floral transition[21], while *JMJ27* regulates flowering genes including *CO* and *FLC*[40]. In many cases, overlapping SNPs showed stronger association signals than di-SNPs or haplotypes. In virus resistance, numerous overlaps clustered near genomic regions previously linked to resistance[24]. In rice, both a SNP and a haplotype were identified near *OsACS6*[31] and *WX1*[32] for AC. These findings suggest a complex interplay among variant types, with single SNPs often capturing part of the overall genetic effect.

Haplotype and di-SNP Overlap and Phenotypic Impact

Haplotypes and di-SNPs overlapped in *Arabidopsis* flowering and virus resistance traits. For example, overlaps occurred near *SAP*[23] (trait DFLD), *NF-YC4*[41] (trait DFLD), and *MYC2*[25] (trait VIR1-RT-qPCR). *NF-YC4* is required for CO-mediated floral promotion[41]. *MYC2* is a conserved viral target. It suppresses terpene-based resistance and promotes vector performance[25]. In rice, a di-SNP and a haplotype were detected near *GW2*[35] for trait PC, and they overlapped. Importantly, haplotype associations with quantitative traits were stronger than those of single constituent SNPs. Characterization of these loci revealed highly significant phenotypic differentiation between haplotypes, whereas phenotypic differences between individual SNP genotypes were generally non-significant. These results underscore the critical role of

haplotypes in genetic analysis in capturing complex genetic effects that are often masked at the single-nucleotide level. Detailed results for all identified loci are provided in Supplementary Tables S1–S3.

Discussion

Conventional GWAS evaluates the association between phenotypes and individual SNPs, a framework that often fails to fully capture the complexity of rare variants and interacting loci. While existing haplotype-based GWAS methods leverage local LD between adjacent SNPs to increase resolution and detection power, they are constrained by using predefined, fixed-length haplotype blocks (e.g., using software like PLINK). Furthermore, because these blocks are typically constructed independently of trait-variant associations, their effectiveness in subsequent GWAS analysis is limited. To address these limitations, we proposed a new method, CHAP-GWAS, which integrates chromosome-scale haplotype analysis with conventional GWAS. In this approach, promising di-SNPs were pre-selected as "seeds" with a loose criterion ($P < 0.05$) and then extended dynamically based on the strength of phenotypic associations. Adopting a less strict threshold in the initial filtering stage captures a broad set of candidate seeds, ensuring that biologically meaningful signals are not prematurely discarded. Conversely, a stringent Bonferroni correction is applied in the second tier of filtering to minimize false positives. This dynamic framework automatically defines and tests haplotype blocks simultaneously, tailoring the process to trait-specific associations. By doing so, CHAP-GWAS fully leverages chromosomal haplotype information and enhances sensitivity to complex genetic variation that is often overlooked by static window-based methods. Additionally, a random-effect model is utilized within CHAP-GWAS to efficiently handle larger numbers of haplotypes, which significantly reduces the computational burden of the fixed-effect models.

Simulation studies demonstrated that CHAP-GWAS consistently outperforms single-SNP, di-SNP, and RAINBOW methods. Our results indicate that detecting 4-SNP haplotypes is more challenging than identifying 3-SNP haplotypes under conditions of low heritability, reflecting the increased complexity of extending seed di-SNPs as haplotype length grows. However, as heritability approaches unity, the performance gap between 3-SNP and 4-SNP analyses diminishes. The robust likelihood ratio test implemented in CHAP-GWAS maintains high statistical power even as haplotype number and complexity increase, making the framework well-suited for the genome-wide analysis of complex traits. In the simulation studies, RAINBOW exhibited much lower detection power in schemes 2 and 3 due to its inherent limitation in differentiating haplotype variants when all constituent SNPs are heterozygous, a common occurrence in our simulated F1 hybrid populations. Nevertheless, this constraint was not observed when analyzing real-world *Arabidopsis* datasets, which consist solely of homozygous accessions. For instance, RAINBOW successfully detected haplotype blocks associated with two flowering traits (4W, LDV) and one virus-related trait (Q-CMV), including a block linked to the *LUH* gene. Furthermore, RAINBOW identified many of the same haplotypes detected by CHAP-GWAS, providing an independent validation of our proposed method.

The complementary value of SNP- and Haplotype-based approaches is further highlighted by the comparative analyses in this study. Certain loci, such as *AGL17*[37], were only identified by single-SNP analysis, likely attributable to the strong effects from individual mutations. In contrast, a haplotype variant was exclusively associated with the gene *LEUNIG* (*LUG*), a transcriptional corepressor that regulates *AGAMOUS* expression during flower development[39]. This demonstrates that multi-SNP contexts can

better capture complex genetic signals. Other loci, including *SVP*[21] and *JMJ27*[40], were identified by both methods. However, single-SNP analysis yielded stronger signals, suggesting dominant individual marker effects. On the contrary, haplotype signals were more significant near *SAP* and *NF-YC4*, likely reflecting regulatory effects more effectively captured by the dynamic extension algorithm of CHAP-GWAS. Genes such as *MYC2* were detected with equal significance by both methods. Together, SNP- and haplotype-based analyses offer a more comprehensive and synergistic view of the underlying genetic architecture.

Analysis of maize datasets further indicated that the performance of CHAP-GWAS, while remaining robust, is sensitive to both sample size and marker density. Reduction in marker density compromises the resolution of genetic variants, as the haplotype growth algorithm (the core of the CHAP-GWAS framework) relies on the local LD structure between adjacent SNPs. Consequently, sparse genotypic data or genomic regions characterized by low coverage and sequencing gaps present challenges for the effective identification and extension of haplotype variants. However, the rapid advancement of high-throughput sequencing and genotyping technologies is expected to mitigate these hurdles in future studies, ensuring broader applicability of the proposed method across diverse genomic landscapes.

A significant advantage of the CHAP-GWAS framework is its extensibility; diverse genomic variants, including complex structural variations (SVs), can be integrated with SNPs to construct composite haplotype variants. This flexibility addresses a substantial limitation of conventional GWAS methodologies and provides a more comprehensive pathway for dissecting the genetic architecture of complex traits. Although CHAP-GWAS captures interactions among adjacent SNPs through haplotype variants, interactions between distant haplotypes were not examined in this study. Conventional GWAS similarly tends to overlook such higher-order interactions. However, CHAP-GWAS can readily be extended to include interaction terms between haplotype variants, which is expected to further enhance its power to detect trait-associated loci characterize their synergy in co-shaping complex traits.

In summary, the development of CHAP-GWAS represents a significant shift toward utilizing chromosome-scale phased data for high-resolution genetic mapping. By transitioning from static, predefined SNP blocks to a dynamic, phenotype-driven extension framework, we have demonstrated a significant improvement in the ability to recover missing heritability and identify regulatory loci that escape traditional single-SNP detection. This method not only offers a robust tool for current genomic datasets but also provides a scalable foundation for future research integrating structural variation and epistatic interactions. As a result, CHAP-GWAS facilitates a more nuanced understanding of how complex genetic architectures collectively influence phenotypic outcomes across diverse biological systems.

Materials and Methods

Linear Mixed Model Incorporating Haplotype Block Random Effects

Let y be a vector for the phenotypic values of n individuals. The genetic model is

$$y = X\beta + Z_k\gamma_k + \sum_{j=1}^{l} W_j a_j + \sum_{j=1}^{l} V_j d_j + \varepsilon \quad (1)$$

where X is an incident matrix for fixed (non-genetic) effects, β is a vector of the fixed effects; Z_k is an $n \times m_k$ matrix of indicators for the kth haplotype block (m_k levels), γ_k is the vector (length of m_k) representing the relative genetic effects of the kth haplotype block with an assumed normal distribution of mean 0 and variance ϕ_k^2; W_j and V_j are the matrices of indicators for additive and dominance effects of the j th marker, respectively; a_j and d_j are the additive and dominance effects for the j th marker with an assumed normal distribution of mean 0 and variance ϕ_a^2 and ϕ_d^2, respectively; l is the number of markers; and ε is a vector of residual errors (length of n) with a multivariate normal distribution with a common residual variance σ^2, *i.e.*, $\varepsilon \square N\left(0, I\sigma^2\right)$. The indicator variable (a vector of length m_k) for the i th individual and k th haplotype block, *i.e.*, $Z_{i,k}$ is defined as

$$Z_{i,k} = \begin{cases} 1,0,0,\ldots,0,0,0 & for \quad haplotype\ level\ 1 \\ 0,1,0,\ldots,0,0,0 & for \quad haplotype\ level\ 2 \\ \vdots & \vdots \\ 0,0,0,\ldots,0,1,0 & for \quad haplotype\ level\ m_k - 1 \\ 0,0,0,\ldots,0,0,1 & for \quad haplotype\ level\ m_k \end{cases} \tag{2}$$

The variables for additive and dominance of individual i at marker j are coded as follows where *A* and *a* represent the two alleles for the j th SNP,

$$W_{i,j} = \begin{cases} 1 & for \quad AA \\ 0 & for \quad Aa \\ -1 & for \quad aa \end{cases} \text{ and } V_{i,j} = \begin{cases} 0 & for \quad AA \\ 1 & for \quad Aa \\ 0 & for \quad aa \end{cases} \tag{3}$$

Two-Step Haplotype Block Testing

Restricted Maximum Likelihood Model with Polygenic Effects

Various methods of inferring a kinship matrix have been proposed. In this study, we used a marker inferred kinship matrix defined as:

$$K_a = \sum_{j=1}^{l} W_j W_j^T \text{ and } K_d = \sum_{j=1}^{l} V_j V_j^T \tag{4}$$

We define a null model as:

$$y = X\beta + \sum_{j=1}^{l} W_j a_j + \sum_{j=1}^{l} V_j d_j + \varepsilon \tag{5}$$

The expectation of y is $E\left(y\right) = X\beta$, and the variance-covariance matrix of y is

$$\text{var}(y) = K_a\phi_a^2 + K_d\phi_d^2 + I\sigma^2 = \left(K_a\lambda_a + K_d\lambda_d + I\right)\sigma^2 = H\sigma^2, \tag{6}$$

where are $\lambda_a = \phi_a^2 / \sigma^2$ and $\lambda_d = \phi_d^2 / \sigma^2$, respectively. After removing the effects of β and σ^2, the restricted maximum likelihood is a function of λ_a and λ_d which is

$$L(\lambda_a, \lambda_d) = -\frac{1}{2}\ln|H| - \frac{1}{2}\ln|X^T H^{-1} X| - \frac{n-r}{2}\ln|y^T P y|, \quad (7)$$

where r is the rank of matrix X and $P = H^{-1} - H^{-1}X\left(X^T H^{-1} X\right)^{-1} X^T H^{-1}$.

Restricted maximum likelihood (REML) solutions of λ_a and λ_d was obtained by maximizing the preceding likelihood function using the Newton-Raphson iterative algorithm. The likelihood was evaluated efficiently using the eigen-decomposition algorithm of Kang *et al*. (2008)[42]. The resulting variance ratio estimates, denoted by $\hat{\lambda}_a$ and $\hat{\lambda}_d$, were then used as constants in the subsequent haplotype block detection model.

Likelihood-Ratio Test for Haplotype Block Detection

The alternative model (equation 1) evaluates the significance of individual haplotype block effects, assuming they are normally distributed with mean zero and a shared variance. The expectation of y remains $E(y) = X\beta$, and the variance-covariance matrix is then

$$\mathrm{var}(y) = Z_k Z_k^T \phi_k^2 + WW^T \phi_a^2 + VV^T \phi_d^2 + I\sigma^2 = Z_k Z_k^T \phi_k^2 + \left(K_a \lambda_a + K_d \lambda_d + I\right)\sigma^2 = Z_k Z_k^T \phi_k^2 + H\sigma^2 \quad (8)$$

where, λ_a and λ_d in H are replaced by the estimated values ($\hat{\lambda}_a$ and $\hat{\lambda}_d$) under the null model (equation 5) and λ_k is defined as $\lambda_k = \phi_k^2 / \sigma^2$. Woodbury matrix identities[43] are applied along with the eigen decomposition to ease the computational burden. Let us perform eigen decomposition for $K_a\hat{\lambda}_a + K_d\hat{\lambda}_d$ so that $K_a\hat{\lambda}_a + K_d\hat{\lambda}_d = UDU^T$, where $D = diag\{\delta_1, \ldots, \delta_n\}$ is a diagonal matrix for the eigenvalues and U is an $n \times n$ matrix for the eigenvectors. Let $y^* = U^T y$, $X^* = U^T X$ and $Z_k^* = U^T Z_k$, be transformed variables so that the restricted log likelihood function for the alternative model (equation 1) is

$$L\left(\lambda_k \middle| \hat{\lambda}_a, \hat{\lambda}_d\right) = -\frac{1}{2}\ln|R_k| - \frac{1}{2}\ln|X^{*T} R_k^{-1} X^*| - \frac{n-r}{2}\ln\left(y^{*T} P_k y^*\right) \quad (9)$$

where r is the rank of matrix X and

$$P_k = R_k^{-1} - R_k^{-1} X^* (X^{*T} R_k^{-1} X^*)^{-1} X^{*T} R_k^{-1} \quad (10)$$

and

$$R_k = Z_k Z_k^T \lambda_k + R \quad (11)$$

where $R = D + I$ is a known diagonal matrix.

Newton's method was again used to estimate λ_k, denoted by $\hat{\lambda}_k$. Under the null hypothesis, $\hat{\lambda}_k = 0$. Thus, the likelihood-ratio test statistic is defined as:

$$\Gamma_k = -2\left[\log L\left(\hat{\lambda}_k \middle| \hat{\lambda}_a, \hat{\lambda}_d\right) - \log L\left(0 \middle| \hat{\lambda}_a, \hat{\lambda}_d\right)\right]. \quad (12)$$

Note that the likelihood function involves $|R_k|$ and R_k^{-1}, both of which are computationally costly. However, the special structure of R_k makes it possible to apply the Woodbury matrix identities[42] to compute $|R_k|$ and R_k^{-1}. Consequently, the random model approach does not lead to a substantial increase in computational time.

P-value of the Test Statistic

Under the null model, this statistic approximately follows a mixture of two chi-squared distributions with equal weights, i.e., $0.5\chi_0^2 + 0.5\chi_1^2$, where χ_0^2 is a fixed number of 0[44,45]. The *p*-value for the haplotype block k, denoted by p_k, it is calculated as:

$$p_k = \begin{cases} 1 & \Gamma_k \le 10^{-5} \\ \frac{1}{2}\Pr\left(\chi_1^2 > \Gamma_k\right) & \Gamma_k > 10^{-5} \end{cases} \quad (13)$$

where Γ_k is the likelihood-ratio test statistic calculated using equation 12, and χ_1^2 is a chi-square random variable with one degree of freedom.

Genome-Wide Haplotype Block Detection

The genome-wide haplotype scanning process of the CHAP-GWAS algorithm, illustrated in Fig. 1b, consists of two steps:

Step 1: Di-SNP Screening. A sliding window strategy was used. The algorithm starts with di-SNP haplotypes generated from adjacent SNPs on each chromosome. For s SNPs on a chromosome, s-1 di-SNP haplotypes were initially tested. Haplotype indicator variables were extracted, classified, and numerically coded for all samples. A linear mixed model, with the di-SNP haplotype as a random effect, assessed the association with the target phenotype. A likelihood-ratio test was used to calculate the P-value. Di-SNPs passing a predefined threshold were designated as seeds for haplotype extension (Step 2), whereas the others were excluded from subsequent analysis.

Step 2: Haplotype Extension. Di-SNP haplotypes passing Step 1 were retained as seeds for haplotype extension. Each seed di-SNP was extended by adding one adjacent SNP to the left and one to the right. Each extended haplotype was then evaluated for association using a linear mixed model. If the better

haplotypes showed a stronger association than the current seed, it became the new seed. Extension then continued in the same direction until the signal no longer improved. The haplotype with the strongest association (lowest P-value) was retained. This process was applied to each seed di-SNP across the genome. The resulting phenotype-associated haplotype blocks included di-SNP variables, extended haplotypes, or both.

Simulation study

A total of 199 *Arabidopsis* lines were randomly paired to generate 200 pseudo-hybrids (Figure 3a). The baseline phenotypic value of each simulated hybrid was calculated as the average value of its selected parents. In the simulated hybrid population, conventional GWAS and CHAP-GWAS were used to identify neutral genomic regions lacking trait-associated SNPs and haplotype blocks, respectively. Subsequently, simulated pseudo trait-associated haplotype blocks were inserted into these neutral regions. This step was used to minimize changes in the trait associations of the original SNPs or haplotype blocks within these regions. As detailed below, three schemes were used to assign the effects of the simulated haplotype blocks on phenotypes.

Scheme 1: We simulated 3-SNP and 4-SNP haplotype blocks, with four and eight alleles, respectively. These simulated blocks replaced the original neutral regions of the same length. For 3-SNP blocks, the artificial genetic effects of the four alleles were assigned as $-3q\sigma_y$, $-q\sigma_y$, $q\sigma_y$ and $3q\sigma_y$, respectively, where σ_y was the standard deviation of the phenotypic values and q was the effect-controlling coefficient with $q = 0.0025, 0.0050, 0.0075, \ldots, 0.1200, 0.1225, 0.1250$, respectively. For 4-SNP blocks, the eight allele effects were assigned as $-7q\sigma_y$, $-5q\sigma_y$, $-3q\sigma_y$, $-q\sigma_y$, $q\sigma_y$, $3q\sigma_y$ $5q\sigma_y$ and $7q\sigma_y$, respectively. For each hybrid, the phenotypic value was calculated as the sum of its baseline value and the simulated effect of the specific haplotype allele. Since the heterozygous genotypes of each constitutional SNPs within the simulated haplotype blocks were identical across all individuals, conventional single-SNP GWAS methods were unable to detect these blocks. Detection power, defined as 1 minus the type II error rate, represented the proportion of tests in which the simulated trait-associated haplotype blocks were correctly identified. The significance threshold was set at 0.05/*m* (Bonferroni correction), where *m* is the total number of SNPs in the original genotype dataset.

Scheme 2: Unlike Scheme 1, we used the original haplotype alleles generated from crossing the 199 *Arabidopsis* lines, rather than simulating new alleles. These natural alleles in the neutral regions had no phenotypic contribution. Artificial genetic effects were assigned to the alleles, with slight variations between odd- and even-numbered alleles at each locus. All other settings were identical to Scheme 1. Detection power for single SNPs was defined as the proportion of tests where at least one trait-associated SNP was successfully identified.

Scheme 3: PLINK was used to select 3-SNP and 4-SNP haplotype blocks from the neutral genomic regions. Artificial genetic effects were then assigned to the various haplotype alleles. All other settings were identical to Schemes 1 and 2.

Public Plant Datasets

Arabidopsis Dataset: We used the homozygous accession dataset originally described by Atwell *et al*., encompassing a diverse panel of natural inbred lines. The initial dataset consisted of 199 samples and 216,130 SNPs[46]. Six flowering time traits were analyzed: long days (DFLD), long days with vernalization (LDV), short days (SD), long days without vernalization (0W), long days with two-week vernalization (2W), and long days with four-week vernalization (4W). After filtering, 216,095 SNPs with no missing data remained. The phenotypic data were log10-transformed to approximate normality.

Additionally, we analyzed an *Arabidopsis* virus response dataset comprising seedlings from 496 and 500 *Arabidopsis* accessions (sourced from the 1001 Genomes Project[47]) infected with CMV-Δ2b and Q-CMV, respectively[24]. Phenotypes were measured as log-transformed ELISA values for Fny-CMV-Δ2b accumulation and RT-qPCR values for Q-CMV accumulation, respectively. After applying a minor allele frequency cutoff of 5%, 1,823,394 and 1,833,408 SNPs were retained for CMV-Δ2b and Q-CMV analyses, respectively.

Maize Agronomy Dataset: We used an NCII-design F1 hybrid population. It comprised 5,820 hybrids derived from 194 maternal and 30 paternal inbred lines representing distinct heterotic groups[48]. Chromosome-scale haplotypes of the F1 hybrids were inferred from parental genotype data. This resulted in a dataset of 2,373,301 SNPs. Plant height (PH) was analyzed as the target trait using 1,000 individuals randomly selected from the 5,820 hybrids.

Rice Agronomy Dataset: We used an F1 hybrid population comprising 565 hybrids derived from 118 parental inbred lines. It was genotyped with 1,619,588 SNPs[49]. Two quality-related traits were analyzed: amylose content (AC) and protein content (PC).

Significance Threshold Selection

For initial haplotype screening, a significance threshold of 0.05 was used for all haplotype-based GWAS runs. For final signal determination, genome-wide significance thresholds were Bonferroni-corrected: $0.05/N_e$ for flowering traits and $1/N_e$ for virus resistance traits, where N_e is the number of effective SNPs. Effective SNP numbers were calculated using Wang *et al*. (2016)[50] for Arabidopsis inbred lines and GEC[51] for rice hybrids.

Gene Annotation and Analysis

Reference genome assemblies TAIR10.1 (NCBI: GCF_000001735.4), B73_v3 (NCBI: GCF_000005005.1), and IRGSP-1.0 (NCBI: GCF_001433935.1) were used for SNP density calculation in Arabidopsis, maize, and rice, respectively.

JBrowse and Gbrowse were used to identify genes near significant haplotype signals. For Arabidopsis data, tair10 (arabidopsis.org) was used; for maize data, MaizeGDB B73 v3 (maizegdb.org) was used; and for rice data, the Rice Annotation Project Database IRGSP1 (rapdb.dna.affrc.go.jp/jbrowse/?data=data/irgsp1) was used.

Data availability

The real Arabidopsis flowering trait data set can be retrieved from: https://www.arabidopsis.org/. The Arabidopsis virus-related phenotype data were obtained from the published study[24], and the genotype data were obtained from the 1001 Genomes Project: https://www.1001genomes.org/. The maize phenotype data were obtained from: https://iagr.genomics.cn/CropGS/#/Datasets, and the maize parental genotype data were obtained from: https://ftp.cngb.org/pub/CNSA/data3/CNP0001565/zeamap/99_MaizegoResources/01_CUBIC_related/Genotype/parent/.

Code availability

The CHAP-GWAS R package was publicly available on R CRAN: https://cran.r-project.org/web/packages/CHAPGWAS/index.html

References:

1. Lee, Sang H., Wray, Naomi R., Goddard, Michael E. & Visscher, Peter M. Estimating Missing Heritability for Disease from Genome-wide Association Studies. *The American Journal of Human Genetics* **88**, 294–305 (2011).
2. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
3. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21 (2008).
4. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569 (2010).
5. Akey, J., Jin, L. & Xiong, M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics* **9**, 291–300 (2001).
6. Zaykin, D. *et al.* Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MGTesting association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53: 79-91. *Human heredity* **53**, 79–91 (2002).
7. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906–913 (2007).
8. Gabriel, S.B. *et al.* The Structure of Haplotype Blocks in the Human Genome. *Science* **296**, 2225–2229 (2002).
9. Browning, B.L., Tian, X., Zhou, Y. & Browning, S.R. Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet* **108**, 1880–1890 (2021).
10. Hofmeister, R.J., Ribeiro, D.M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nature Genetics* **55**, 1243–1249 (2023).
11. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics* **48**, 1443–1448 (2016).

12. Bhat, J.A., Yu, D., Bohra, A., Ganie, S.A. & Varshney, R.K. Features and applications of haplotypes in crop breeding. *Communications Biology* **4**, 1266 (2021).
13. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).
14. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
15. Qi, G. *et al.* Reveal genomic insights into cotton domestication and improvement using gene level functional haplotype-based GWAS. *Nature Communications* **16**, 4734 (2025).
16. Han, X. *et al.* Two haplotype-resolved, gap-free genome assemblies of Actinidia latifolia and Actinidia chinensis shed light on regulation mechanisms of vitamin C and sucrose metabolism in kiwifruit. *Molecular Plant* **16**(2022).
17. Hou, Y. *et al.* Haplotype-based pangenomes reveal genetic variations and climate adaptations in moso bamboo populations. *Nature Communications* **15**, 8085 (2024).
18. Huang, J. *et al.* Haplotype-resolved gapless genome and chromosome segment substitution lines facilitate gene identification in wild rice. *Nature Communications* **15**, 4573 (2024).
19. Hamazaki, K. & Iwata, H. RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. *PLOS Computational Biology* **16**, e1007663 (2020).
20. Li, R. *et al.* Inference of Chromosome-Length Haplotypes Using Genomic Data of Three or a Few More Single Gametes. *Molecular Biology and Evolution* **37**, 3684–3698 (2020).
21. Hartmann, U. *et al.* Molecular cloning of SVP: a negative regulator of the floral transition in Arabidopsis. *The Plant Journal* **21**, 351–360 (2000).
22. Sitaraman, J., Bui, M. & Liu, Z. LEUNIG_HOMOLOG and LEUNIG Perform Partially Redundant Functions during Arabidopsis Embryo and Floral Development *Plant Physiology* **147**, 672–681 (2008).
23. Byzova, M. *et al.* Arabidopsis STERILE APETALA, a multifunctional gene regulating inflorescence, flower, and ovule development. *Genes & Development* **13**, 1002–1014 (1999).
24. Liu, S. *et al.* Identification of positive and negative regulators of antiviral RNA interference in Arabidopsis thaliana. *Nature Communications* **13**, 2994 (2022).
25. Li, R. *et al.* Virulence factors of geminivirus interact with MYC2 to subvert plant resistance and promote vector performance. *Plant Cell* **26**, 4991–5008 (2014).
26. Xie, B. *et al.* NUA positively regulates plant immunity by coordination with ESD4 to deSUMOylate TPR1 in Arabidopsis. *New Phytologist* **241**, 363–377 (2024).
27. Ishihara, T. *et al.* Overexpression of the Arabidopsis thaliana EDS5 gene enhances resistance to viruses. *Plant Biology* **10**, 451–461 (2008).
28. Li, L. *et al.* Joint analysis of phenotype-effect-generation identifies loci associated with grain quality traits in rice hybrids. *Nature Communications* **14**, 3930 (2023).
29. Guo, N., Gu, M., Hu, J., Qu, H. & Xu, G. Rice OsLHT1 Functions in Leaf-to-Panicle Nitrogen Allocation for Grain Yield and Quality. *Frontiers in Plant Science* **Volume 11 - 2020**(2020).
30. Du, Y. *et al.* Flowering-Related RING Protein 1 (FRRP1) Regulates Flowering Time and Yield Potential by Affecting Histone H2B Monoubiquitination in Rice (Oryza Sativa). *PLOS ONE* **11**, e0150458 (2016).
31. Matsushima, R. *et al.* Amyloplast Membrane Protein SUBSTANDARD STARCH GRAIN6 Controls Starch Grain Size in Rice Endosperm. *Plant Physiology* **170**, 1445–1459 (2016).
32. Sano, Y. Differential regulation of waxy gene expression in rice endosperm. *Theoretical and Applied Genetics* **68**, 467–473 (1984).
33. Fan, C. *et al.* GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theoretical and Applied Genetics* **112**, 1164–1171 (2006).

34. Miao, J. *et al.* Mutation of RGG2, which encodes a type B heterotrimeric G protein γ subunit, increases grain size and yield production in rice. *Plant Biotechnology Journal* **17**, 650–664 (2019).
35. Song, X.-J., Huang, W., Shi, M., Zhu, M.-Z. & Lin, H.-X. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nature Genetics* **39**, 623–630 (2007).
36. Michaels, S.D. & Amasino, R.M. FLOWERING LOCUS C Encodes a Novel MADS Domain Protein That Acts as a Repressor of Flowering. *The Plant Cell* **11**, 949–956 (1999).
37. Han, P., García-Ponce, B., Fonseca-Salazar, G., Alvarez-Buylla, E.R. & Yu, H. AGAMOUS-LIKE 17, a novel flowering promoter, acts in a FT-independent photoperiod pathway. *The Plant Journal* **55**, 253–265 (2008).
38. Zhao, Y. *et al.* HANABA TARANU Is a GATA Transcription Factor That Regulates Shoot Apical Meristem and Flower Development in Arabidopsis. *The Plant Cell* **16**, 2586–2600 (2004).
39. Conner, J. & Liu, Z. LEUNIG, a putative transcriptional corepressor that regulates AGAMOUS expression during flower development. *Proceedings of the National Academy of Sciences* **97**, 12902–12907 (2000).
40. Dutta, A., Choudhary, P., Caruana, J. & Raina, R. JMJ27, an Arabidopsis H3K9 histone demethylase, modulates defense against Pseudomonas syringae and flowering time. *The Plant Journal* **91**, 1015–1028 (2017).
41. Kumimoto, R.W., Zhang, Y., Siefers, N. & Holt Iii, B.F. NF–YC3, NF–YC4 and NF–YC9 are required for CONSTANS-mediated, photoperiod-dependent flowering in Arabidopsis thaliana. *The Plant Journal* **63**, 379–391 (2010).
42. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
43. Golub, G.H. & Van Loan, C.F. *Matrix Computations - 4th Edition*, (Johns Hopkins University Press, Philadelphia, PA, 2013).
44. Chernoff, H. On the Distribution of the Likelihood Ratio. *The Annals of Mathematical Statistics* **25**, 573–578 (1954).
45. Visscher, P.M. A Note on the Asymptotic Distribution of Likelihood Ratio Tests to Test Variance Components. *Twin Research and Human Genetics* **9**, 490 – 495 (2006).
46. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**, 627–31 (2010).
47. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).
48. Yang, W. *et al.* Target-oriented prioritization: targeted selection strategy by integrating organismal and molecular traits through predictive analytics in breeding. *Genome Biology* **23**, 80 (2022).
49. Wang, S.-B. *et al.* Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific Reports* **6**, 19444 (2016).
50. Wang, Q., Wei, J., Pan, Y. & Xu, S. An efficient empirical Bayes method for genomewide association studies. *Journal of Animal Breeding and Genetics* **133**, 253–263 (2016).
51. Li, M.-X., Yeung, J., Cherny, S. & Sham, P. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human genetics* **131**, 747–56 (2011).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalTableS1.pdf
- SupplementalTableS2.pdf
- SupplementalTableS3.pdf
- SupplementalFigures.pdf