bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Causal Haplotype Block Identification in Plant Genome-Wide Association Studies

Xing Wu[1+], Wei Jiang[2+], Chris Fragoso[1], Jing Huang[3], Geyu Zhou[4], Hongyu Zhao[2,4], Stephen Dellaporta[1*]

**Affiliation:**

[1] Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut, USA

[2] Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, Connecticut, USA

[3] Department of Applied Biological Science, Zhejiang University, Hangzhou, Zhejiang, China

[4] Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

[*] Corresponding author. E-mail: stephen.dellaporta@yale.edu

[+] These authors contribute equally to this work.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.


## Abstract

Genome wide association studies (GWAS) can play an essential role in understanding genetic basis of complex traits in plants and animals. Conventional SNP-based linear mixed models (LMM) used in many GWAS that marginally test single nucleotide polymorphisms (SNPs) have successfully identified many loci with major and minor effects. In plants, the relatively small population size in GWAS and the high genetic diversity found many plant species can impede mapping efforts on complex traits. Here we present a novel haplotype-based trait fine-mapping framework, HapFM, to supplement current GWAS methods. HapFM uses genotype data to partition the genome into haplotype blocks, identifies haplotype clusters within each block, and then performs genome-wide haplotype fine-mapping to infer the causal haplotype blocks of trait. We benchmarked HapFM, GEMMA, BSLMM, and GMMAT in both simulation and real plant GWAS datasets. HapFM consistently resulted in higher mapping power than the other GWAS methods in simulations with high polygenicity. Moreover, it resulted in higher mapping resolution, especially in regions of high LD, by identifying small causal blocks in the larger haplotype block. In the Arabidopsis flowering time (FT10) datasets, HapFM identified four novel loci compared to GEMMA's results, and its average mapping interval of HapFM was 9.6 times smaller than that of GEMMA. In conclusion, HapFM is tailored for plant GWAS to result in high mapping power on complex traits and improved mapping resolution to facilitate crop improvement.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## Introduction

Genome-wide association study (GWAS) presents a powerful tool to link genetic variations with phenotypic traits. In human studies, GWAS has been extensively employed to associate numerous genetic variants with candidate genes responsible for human diseases, some of which have become targets for medical interventions [1]. For example, the identification of an androgen receptor (AR) gene through GWAS led to the development of therapeutic drugs for patients with prostate cancer [2]. GWAS methods have also been used in plant studies to identify the genetic basis of certain agronomic traits (reviewed by [3]). There have been many successful applications including the identification of OsSPY for plant architecture in rice [4], metabolic genes for tomato flavor [5], and ZmFBL41 for blight resistance in maize [6]. Although genetic associations in plants have been revealed through GWAS, serious limitations still exist in the current best practices, including insufficient power and poor biological interpretation [3,7 8,9]. For the most part, these limitations are due to the relatively small population size in plant studies, usually in the hundreds, reducing mapping power as compared to human GWAS analyses that may involve tens of thousands of individuals.

Mapping power is critical for understanding the genetic architecture of complex traits in GWAS. Many agronomic traits, such as yield, flowering time and disease resistance, are complex in nature involving many loci with variable effect sizes, some of which are difficult to be identified due to systemic issues in most plant GWAS datasets: small population size, existing confounding factors such as population structure and kinship between individuals, and a high levels of genetic diversity common to plant genomes [3,8]. Conventional SNP-based GWAS methods use linear mixed models (LMM) to account for population structure and kinship and then marginally regress individual variants to test for significance. A few variations of the LMM-based methods such as MLMM [10], SUPER [11] and FarmCPU [12] have been proposed to increase mapping power. These GWAS models, however, still have insufficient power because true causal variants may have small effects, and the models lack power to detect minor effect loci because of the small population size. Moreover, a large number of variants causes multiple testing burden further reducing detection power [3]. In human GWAS studies, SNP-set based GWAS

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

method, SMMAT [13] has been proposed to increase the mapping power by grouping nearby variants to aggregate small effects to reduce the number of tests. This method has yet to be evaluated in plant mapping studies. In the recent years, haplotype-based GWAS methods, such as RAINBOW [14] and FH-GWAS [15], were developed which showed improvements in mapping power over SNP-based methods in plant datasets. These studies have demonstrated the feasibility of using haplotypes as variables to overcome issues in plant GWAS.

In addition to mapping power, mapping resolution is another critical aspect of GWAS with small mapping intervals benefitting downstream experimental validation. Many plant species, especially those propagated via self-pollinating or vegetative cloning, have extensive LD block structures [16-18]. For a significant locus in the high LD region, conventional GWAS methods identify variants with significant *p*-values without differentiating causal from proximal variants. This can result in a large mapping interval spanning over dozens or hundreds of genes [3 19], greatly increasing the difficulty of downstream validations.

A typical approach to increasing mapping resolution in plant mapping studies is to generate fine-mapping populations to enhance recombination in the targeted region [20-22]. This approach, however, is an escalation in time, sometimes years, and effort and an option that is not always feasible. Post GWAS analyses such as statistical fine-mapping models have been proposed in human genetics, which can leverage biological annotations to identify potential causal variants among linked genetic variants [23]. These methods, however, restrict fine-mapping analyses to significant GWAS loci only, which limits their utility in plant studies. Similar to SNP-set based association methods, statistical fine-mapping methods have not been adequately evaluated in plant studies yet.

As a result of the rapid growth in sequence-based resources, many plant species now, or in the near future, have extensive genomic resources available to complement the study of genetic basis of complex traits. In plants, complex variations, such as structural variation (SVs), are often the drivers of many quantitative traits, and genome-wide catalogs of SVs are fast becoming available for many plant species, including Arabidopsis [24], rice [25], tomato [26], soybean [27], maize [28] to name a few. Similarly, the

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

availability of transcriptomic datasets can be utilized to identify gene expression changes that result in phenotypic alteration in plants [29]. Yet, in the past, conventional plant GWAS methods have not been capable of incorporating these resources into the trait mapping pipeline. Therefore, a novel trait mapping framework that can systemically incorporate informative genomic, transcriptomic and other meta-datasets to increase mapping power would represent a significant improvement over current methodologies.

In this paper, we present a novel haplotype-based trait fine mapping framework, HapFM, that addresses limitations in plant GWAS methodologies. Unlike previous haplotype-based mapping algorithms, HapFM incorporates the use of unique haplotypes clusters based on historical recombination, rather than individual SNPs or uniform block partitioning of SNPs, to fit a genome-wide statistical fine-mapping model. Furthermore, HapFM was designed to permit the systemically incorporate biological annotations such as SV and other biological elements to facilitate causal inference and biological interpretation of the mapping results. Compared to previous GWAS methods, HapFM resulted greater mapping power and smaller mapping intervals for complex traits with both simulated and real plant datasets. In addition, we demonstrated that it is possible to incorporate SV and functional annotation datasets into HapFM to further increase mapping power. Overall, HapFM achieves a balance between statistical power interpretability, and downstream experimental verifiability.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## Results

*Overview of HapFM workflow*

In this paper, we present a novel haplotype-based trait fine-mapping framework, HapFM, to serve as a powerful strategy for mapping complex traits(Figure 1). There are four steps in the HapFM framework: block partition, unique haplotype identification, haplotype clustering, and statistical fine mapping. In the block partition step, HapFM identifies genome-wide haplotype blocks based on LD information. In order to increase computation efficiency, HapFM utilizes a 2-step partitioning strategy. It first identifies large independent blocks which are defined as a set of adjacent SNPs with minimum pairwise LD ($r^2$) greater than a pre-defined threshold ($r^2 = 0.1$ by default). Next, HapFM partitions each independent block into sub-blocks using available block partition programs. The block partition step outputs non-overlapping SNP sets representing haplotype blocks in the genome.

In the haplotype identification step, HapFM enumerates a set of unique haplotypes in each block based on phased SNP genotypes. If the number of unique haplotypes exceeds the user-defined threshold (n = 10 by default), HapFM will cluster unique haplotypes to reduce the number of variables in the mapping step. After the haplotype clustering step, HapFM outputs a haplotype design matrix which will be used for statistical fine mapping. The haplotype design matrix also has the same format as the conventional SNP genotype matrix, therefore it is compatible to current GWAS methods as well.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

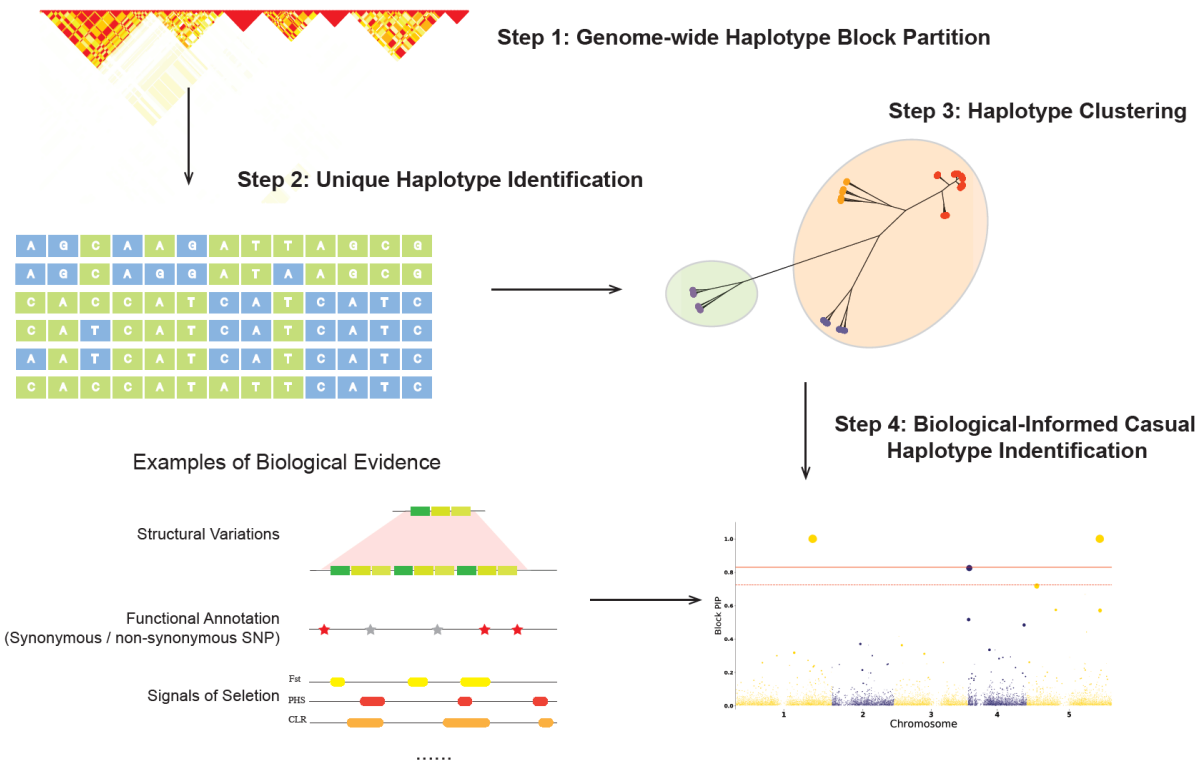



***Figure 1.*** The workflow of haplotype-based trait fine mapping (HapFM). HapFM consists of four steps: genome-wide haplotype block partition, unique haplotype identification, haplotype clustering, and causal haplotype identification. Biological features, such as structural variations, functional annotations, signals of selection, etc. can be incorporated into the fine mapping model. The *y*-axis of Manhattan plot generated by HapFM is block pip, indicating causal probability. The size of the dots indicates the effect size of the block.

In the genome-wide statistical fine mapping step, HapFM follows a linear mixed model (LMM) and a hierarchical Bayes inference framework to infer the causal relationship between haplotype blocks and the phenotype. Upon availability, HapFM can also incorporate existing biological evidence to model the prior probability of causality for each haplotype block. The fine-mapping model accounts for the LD between haplotype blocks, and therefore the result suggests the causal instead of association relationship with the phenotype.

*Block partition and haplotype clustering algorithms*

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Various algorithms were benchmarked to assess the robustness of block partitioning and haplotype clustering steps used in HapFM. Four clustering methods: affinity propagation [30], X-means [31], KNN-spectral clustering and local-spectral clustering [32], were first benchmarked for the clustering step. A high haplotype diversity dataset was simulated to contain, on average, 500 blocks and 15 unique haplotypes derived from three founder haplotypes in each block. Both low and high polygenicity trait datasets were tested for comparative purposes. Comparable mapping power was found for the low polygenicity simulations and none of the clustering methods consistently outperformed the others (Figure 2a, Supplemental Figure 1a). In the high polygenicity datasets, affinity propagation and X-means clustering methods consistently resulted in higher mapping power than KNN-spectral and local-spectral clustering (Supplemental Figure 1b). Different clustering algorithms resulted in similar true positive rate in both low and high polygenicity simulations (Supplemental Figure 2). Affinity propagation gave 2.7 times more clusters than X-means in real data analyses, which costs longer computational time in the mapping step (Supplemental Table 1). Overall, considering user-friendliness, mapping power, and computational time, X-means was found to be more favorable than the other three cluster methods tested.

Next, we compared three different block partition algorithms -- BigLD, Plink, and a uniform partition method -- with the simulated ground truth for block partition accuracy. BigLD and Plink generated outputs closer to the true partitions in the low haplotype diversity setting while BigLD outperformed Plink when analyzing high diversity simulations, whose genome partitions were numerous small blocks that failed to capture local LD structures (Supplemental Figure 3). Uniform partitioning underperformed in both datasets suggesting that the fixed size of blocks was a poor reflection of the underlying LD structure.

We then compared the trait mapping power using haplotype blocks identified by each method in simulated datasets. The simulated datasets covered both low and high haplotype diversity and trait polygenicity, and four types of QTL architectures which represented different numbers of major and minor effect alleles in each locus (Figure 2a). Minor mapping power differences were found between BigLD and Plink blocks in the low haplotype diversity simulations. When the trait polygenicity was low,

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

BigLD blocks consistently resulted in higher or comparable mapping power than that of Plink blocks in all four QTL architectures (Supplemental Figure 4a). When the trait polygenicity was high, BigLD blocks resulted in higher mapping power than that of Plink blocks in QTL architecture 1 and 3 scenarios, and comparable mapping power in the QTL architecture No. 2 and 4 scenarios (Supplemental Figure 4b). The mapping power of BigLD blocks was similar to ground truth blocks, and uniformed partition blocks had the lowest mapping power consistently.

Major mapping power differences were found between BigLD and Plink blocks in the high haplotype diversity simulations. BigLD blocks consistently resulted in higher mapping power than that of Plink blocks in all four QTL scenarios in both low and high polygenicity simulations (Figure 2b, 2c). Plink blocks resulted in similar mapping power as that of uniform partitions.

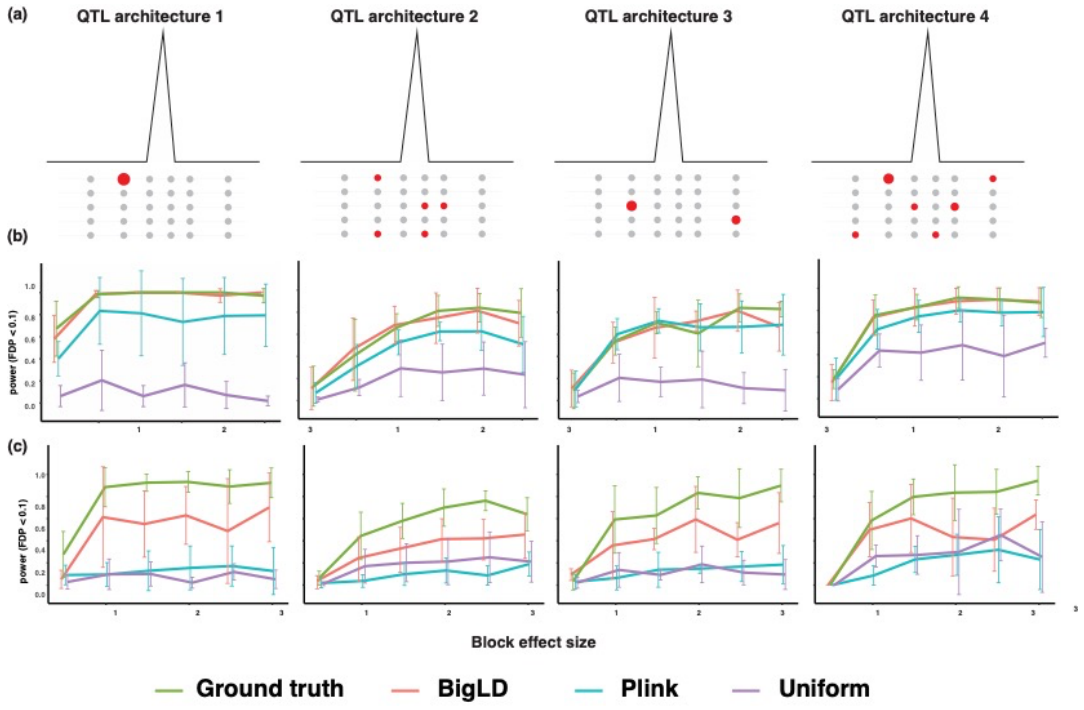



***Figure 2.*** Simulation schemes and mapping power comparison of different block partition algorithms.

(a) Four types of QTLs simulated in the datasets. The effect of QTL1 is contributed by one large effect SNP. The effect of QTL2 is contributed by several minor effect SNPs which are not on

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the same haplotypes. The effect of QTL3 is contributed by two modest effect SNPs which are not on the same haplotype. The effect of QTL4 is contributed by a mixture of modest and small effect SNPs that are not on the same haplotypes. (b) Mapping power comparison (FDR < 0.05) of block partition algorithms in the low haplotype diversity and high polygenicity simulations. The *x*-axis indicates the per-locus heritability. (c) Mapping power comparison (FDR < 0.05) of block partition algorithms in the high haplotype diversity and high polygenicity simulations. The *x*-axis indicates the per-locus heritability.

*GWAS algorithms on simulated datasets*

Four GWAS algorithms: GEMMA, HapFM, BSLMM, and GMMAT, were studied for true positive rate, mapping power, and interval length in simulated datasets. When the trait polygenicity and haplotype diversity were both low, GEMMA consistently gave the highest mapping power and smallest standard deviation in the low haplotype diversity simulations. HapFM and GMMAT provided comparable mapping power to GEMMA in QTL architecture 2, and both HapFM and GMMAT displayed similar mapping power in all four QTL architectures. BSLMM consistently resulted in the lowest mapping power (Supplemental Figure 5a). GEMMA, HapFM, and GMMAT resulted in similar true positive rates, which were significantly higher than that of BSLMM (Supplemental Figure 6a).

When the trait polygenicity was low and haplotype diversity was high, GEMMA resulted in the highest mapping power and smallest standard deviation in QTL architectures 1, 3, and 4. HapFM resulted in similar mapping power to GEMMA in QTL architecture 2 and HapFM consistently resulted in higher or similar mapping power than GMMAT in four QTL scenarios. BSLMM consistently resulted in the lowest mapping power, but its mapping power was increased in the high diversity simulations compared to the low haplotype diversity simulations (Supplemental Figure 5b). HapFM resulted in higher true positive rate than GEMMA and GMMAT in QTL architecture 1, and the true positive rates of the three were comparable in QTL architectures 2, 3, and 4.

When the trait polygenicity was high, HapFM consistently resulted in the highest mapping power in all four QTL architectures in both low and high haplotype diversity simulations (Figure 3). As

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

expected, the mapping power of HapFM decreased in the low diversity simulations. The true positive rate of HapFM was consistently higher than or similar to those of GEMMA, GMMAT, and BSLMM (Supplemental Figure 7).

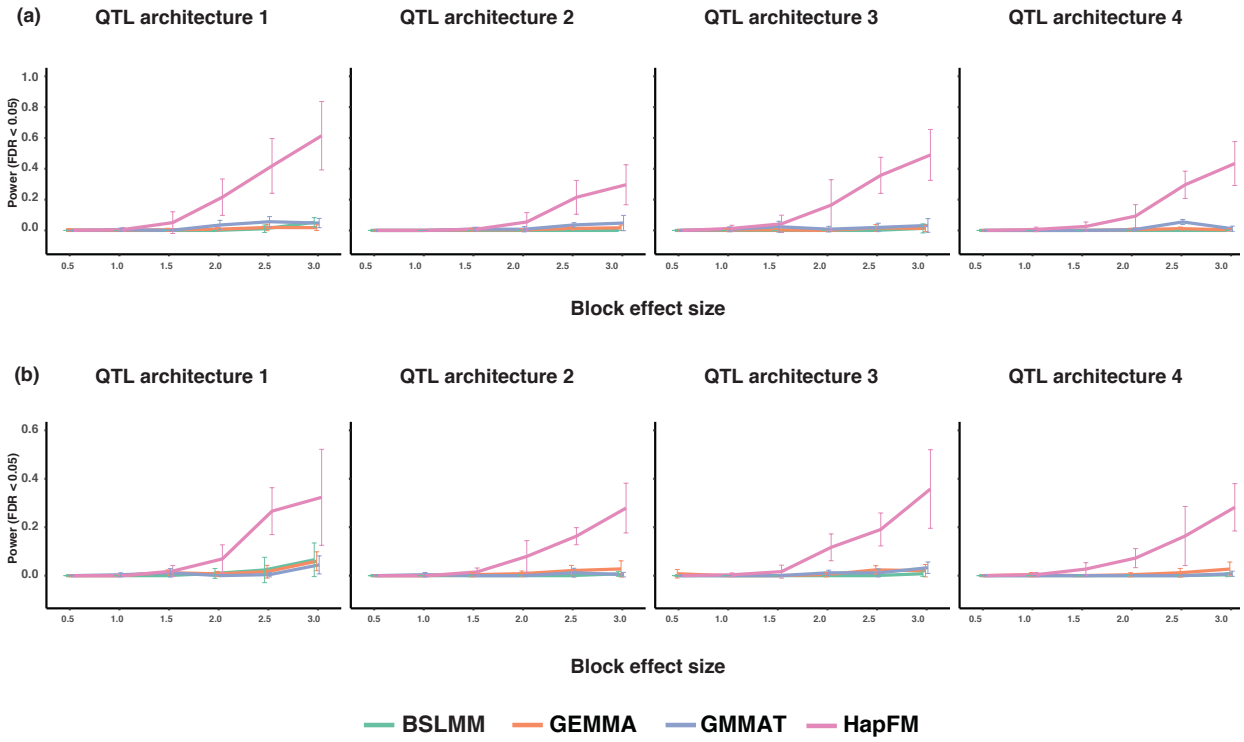



***Figure 3.*** Mapping power comparisons of different GWAS algorithms in the high polygenicity simulations. The x-axis indicates the per-locus heritability.
(a) Mapping power comparisons (FDR < 0.05) of different GWAS algorithms in the low haplotype diversity and high polygenicity simulations. (b) Mapping power comparisons (FDR < 0.05) of different GWAS algorithms in the high haplotype diversity and high polygenicity simulations.

The mapping interval length of significant loci of GEMMA resulted in higher variation than those of HapFM, BSLMM, and GMMAT in all trait polygenicity and haplotype diversity simulations. When the trait polygenicity was low, the average interval length of GEMMA significant loci was 29.53 times higher than that of HapFM in the low haplotype diversity simulation. Similarly, the average interval length of GEMMA significant loci was 23.32 times higher than that of HapFM (Figure 4a) in the high haplotype diversity simulation. When the trait polygenicity was high, the average interval length of GEMMA

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

significant loci was 15.19 times higher than that of HapFM in the low haplotype diversity simulations. The average interval length of GEMMA significant loci was 13.32 times higher than that of HapFM in the high haplotype diversity simulations (Figure 4b). The median interval length of GEMMA was not significantly different from that of HapFM (median test, p-value 0.37). In addition, the variance of the interval length of significant loci of GEMMA was significantly higher than those of the other three GWAS algorithms in all the simulations (Supplemental Table 1).

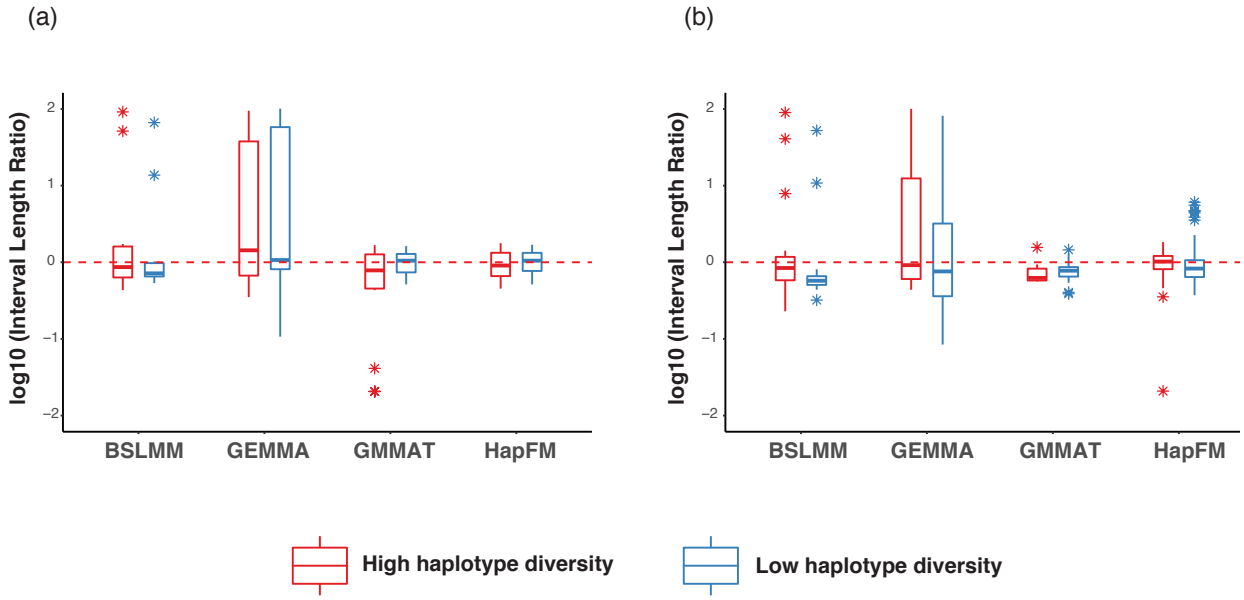



***Figure 4.*** Mapping interval comparisons of different GWAS algorithms in the simulations. The interval length ratio was calculated by normalizing to the average HapFM's interval length. The red dash line indicates the average interval length of significant signals identified by HapFM.

(a). Interval length of significant loci (FDR < 0.05) identified by different GWAS algorithms in the low polygenicity simulations. (b). Interval length of significant loci (FDR < 0.05) identified by different GWAS algorithms in the high polygenicity simulations

*GWAS algorithms on actual plant datasets*

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Five plant GWAS datasets -- Arabidopsis flower time, rice heading time, cassava HCN content, tomato metabolite concentration, and maize height -- were used to benchmark the performance of HapFM as compared to the other GWAS algorithms (Table 1). HapFM identified the most significant loci compared to the other GWAS algorithms in the Arabidopsis flowering time (FT10) dataset (Figure 5). HapFM first partitioned genome into 48,171 haplotype blocks, out of which it identified 82,431 haplotype clusters. The average and median of block length were 2,803 nt and 457 nt, respectively. In the haplotype fine mapping step, HapFM identified seven significant loci (FDR < 0.05). GEMMA identified five significant loci (FDR < 0.05), out of which three loci were shared with HapFM results. The locus on Chr5 (most significant SNP: 5@3161477) was also detected by HapFM but slightly missed the significant FDR cutoff (FDR = 0.07). GMMAT identified two significant loci and both of them were identified as significant by HapFM and GEMMA. BSLMM identified one significant locus also discovered by HapFM and GEMMA. HapFM identified four loci: Chr3@7598564-7598957, Chr4@405136-406621, Chr5@14063228-14197451, and Chr5@16141604-16146257 that were unique to HapFM algorithm. In these unique intervals, flowering time related candidate genes were identified in or near those loci. In the Chr3@7598564-7598957 locus, there is no gene in the interval but an adjacent proximal gene AT3G21570 located 1.3kb away, was previously shown to be exclusively expressed in the developing flowers with transcriptomic changes during pollen germination and tube growth in Arabidopsis [33]. The Chr4@405136-40662 interval overlaps with AT4G00950 (MEE47), a gene that is highly expressed in mature flowers and required for female gametophyte development and function in Arabidopsis [34 35]. In the Chr5@14063228-14197451 interval, there are 30 protein-coding genes. Multiple candidate genes in the interval, such as AT5G36110, AT5G35926, AT5G35995, have been shown to be highly expressed in different flower stages and tissues [36]. The Chr5@16141604-16146257 locus overlaps with AT5G40360 (MYB115), a gene was shown to be highly expressed during flowering stages and mature flowers and its overexpression promotes vegetative-to-embryonic transition in Arabidopsis [37].

In addition to having the highest mapping power, HapFM also mapped significant loci to the smallest genomic intervals in most cases. For example, HapFM, GEMMA, and BSLMM all identified the same

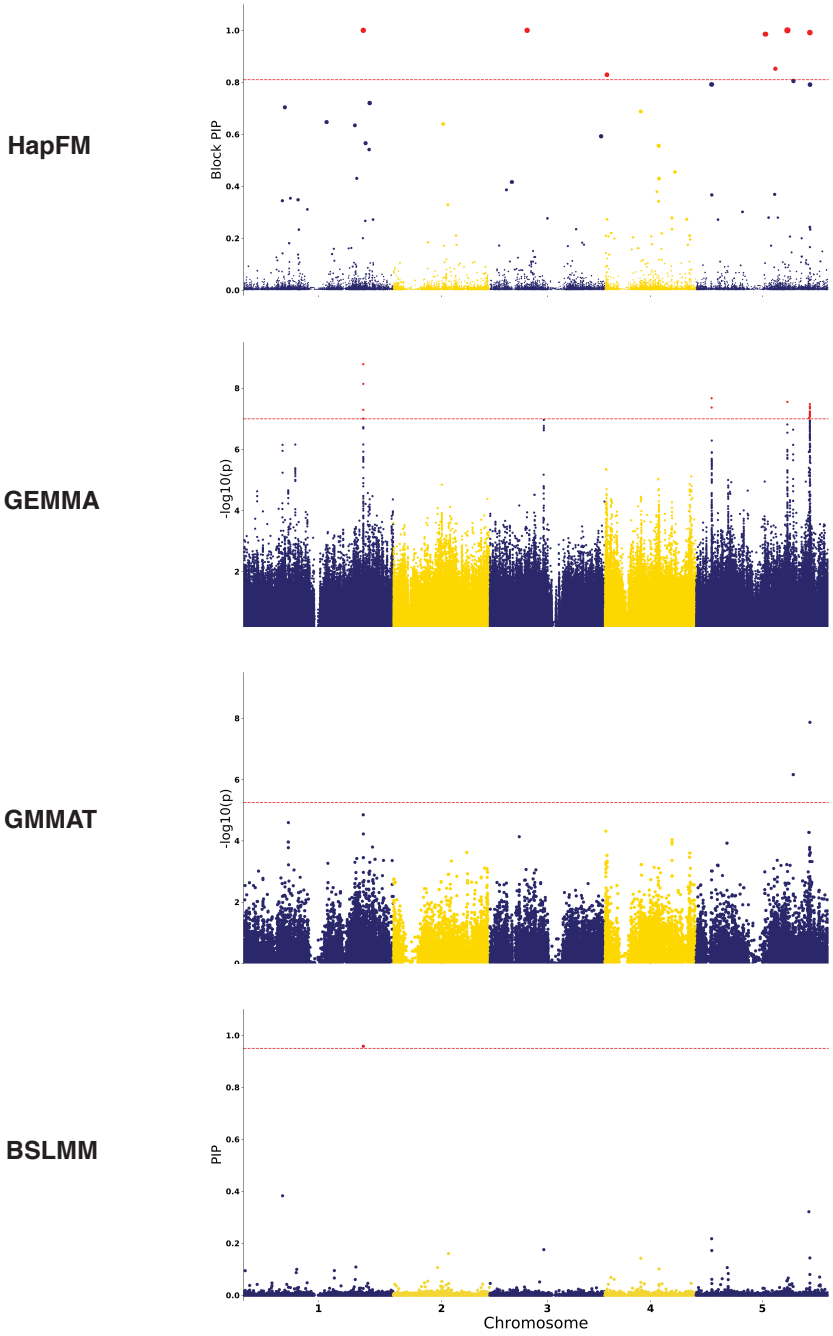
bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

significant locus, FT locus, on Chromosome 1 (Figure 5). The interval length of the locus identified by GEMMA and BSLMM are both 21.9kb while the interval length of the locus identified by HapFM is 2.7kb. On average, the average interval length of significant loci identified by HapFM and GEMMA was 24.8kb and 237.8kb, respectively (Table 1). The average number of SNPs per significant locus identified by HapFM and GEMMA was 28 and 105, respectively. Similar results were found in the other four real plant GWAS datasets (Table 1). HapFM consistently resulted in similar or higher number of significant loci than GEMMA, BSLMM, and GMMAT. In addition, the mapping interval of HapFM is considerably smaller than GEMMA in all the comparisons.

Using the Arabidopsis flowering time dataset, a proof-of-concept study demonstrated that biological annotations could be incorporated (HapFM-anno) and potentially increase mapping power. The biological-informed prior probability for each haplotype block was calculated using eight biological annotations. In this example, the biological annotations were the number of CNV, INDEL, rare variants, high effect variants, moderate effect variants, low effect variants, and modifier variants in each block. The estimated effect size of biological annotations suggested the number of CNV in each block significantly affected the prior probability of each haplotype block (Figure 6a). HapFM-anno identified nine significant loci in total using biological-informed priors (Figure 6b,c). Five out of nine were also identified previously without biological annotation incorporated. HapFM-anno identified four novel loci: Chr1@7884994-7886542, Chr1@11474330-11475120, Chr1@25408933-25429985, and Chr5@23204856-23205070 (Figure 6b). The interval Chr1@7884994-7886542 is at the upstream region of gene AT1G22330 that is highly expressed in mature flowers [36]. The interval Chr1@11474330-11475120 is at the upstream of the gene AT1G31940 that is highly expressed in mature flowers [36] and involved in seed germination [38]. The locus Chr1@25408933-25429985 overlaps with ten genes. Multiple candidate genes in the interval, such as AT1G67780 and AT1G67790, have been shown to be highly expressed during petal differentiation and expansion stage [36]. The locus Chr5@23204856-23205070 overlaps with the gene AT5G57280 that has been shown to be highly expressed in different flower tissues [36] and pre-meristematic cell-mound formation during shoot regeneration [39]. Two HapFM identified loci:

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Chr5@14063228-14197451 and Chr5@16141604-16146257, were not significant after incorporating biological annotations.



bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

***Figure 5.*** Manhattan plots of different GWAS methods on the Arabidopsis flowering time (FT10) dataset. The red dash line indicates the FDR 0.05 threshold. In the HapFM's plot, the size of the dots indicates the estimated effect size of the block.

| Phenotype | Dataset | Block number | GWAS algorithms | # of significant loci (FDR < 0.05) | Avg. significant locus length (nt) | Avg. # of snps per locus |
|---|---|---|---|---|---|---|
| Arabidopsis Flowering | 1003 individuals 1.12M SNPs | 48,171 | HapFM | 7 | 24,780 | 28 |
| | | | GEMMA | 6 | 237,772 | 105 |
| | | | BSLMM | 1 | 21,863 | 80 |
| | | | GMMAT | 2 | 10,110 | 27 |
| Rice Heading Time | 529 individuals 1.43M SNPs | 14,301 | HapFM | 20 | 236,200 | 63 |
| | | | GEMMA | 10 | 2,024,412 | 517 |
| | | | BSLMM | 1 | 53,189 | 43 |
| | | | GMMAT | 4 | 249,122 | 66 |
| Cassava HCN | 1134 individuals 24.75K SNPs | 9,112 | HapFM | 3 | 62,018 | 44 |
| | | | GEMMA | 4 | 1,068,992 | 348 |
| | | | BSLMM | 0 | NA | NA |
| | | | GMMAT | 2 | 71,044 | 32 |
| Maize Height | 263 individuals 23.09M SNPs | 98,723 | HapFM | 10 | 398,161 | 62 |
| | | | GEMMA | 0 | NA | NA |
| | | | BSLMM | 0 | NA | NA |
| | | | GMMAT | 2 | 312,091 | 70 |

Table 1. Summary of GWAS results on the five real plant datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

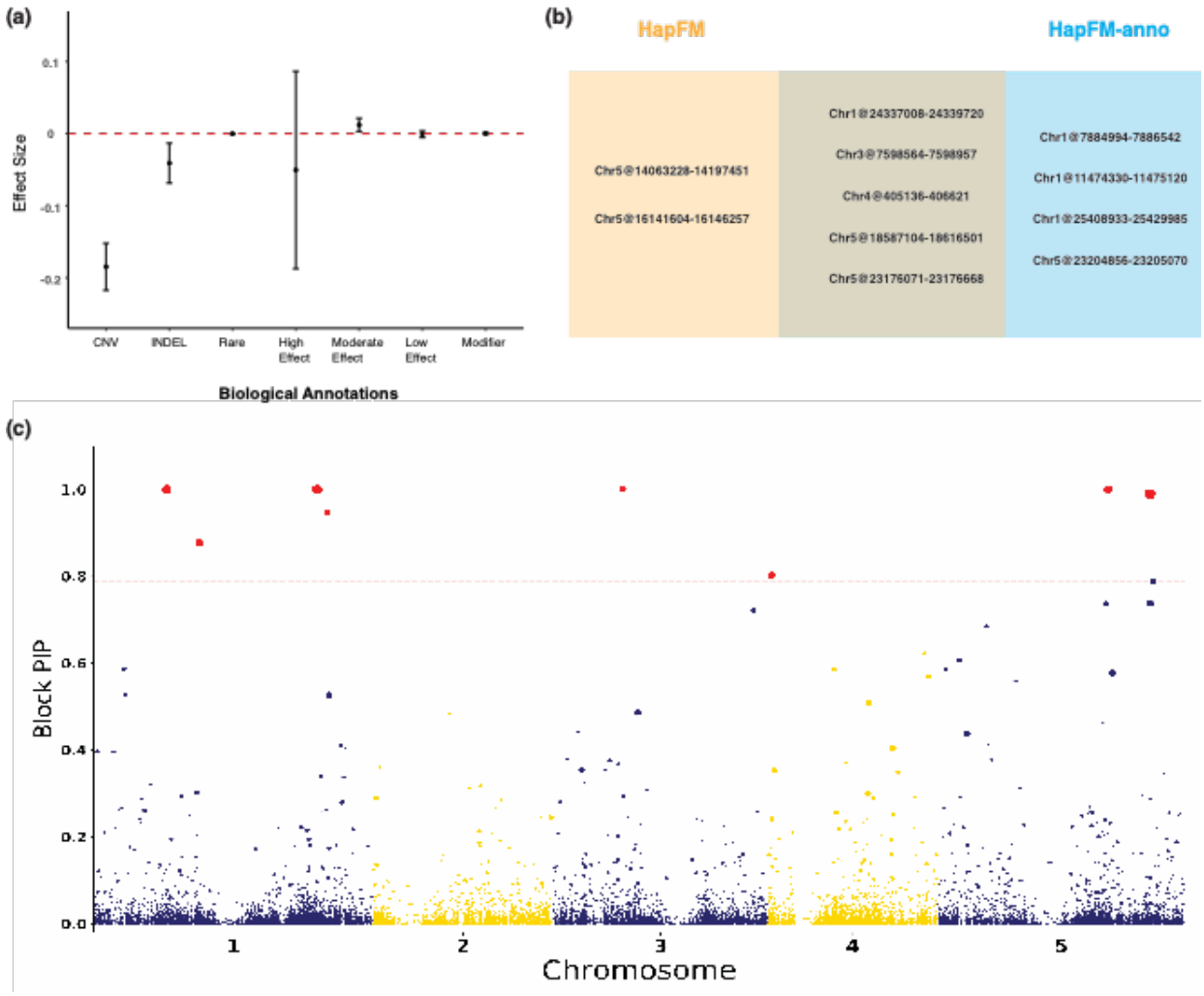



***Figure 6***. Arabidopsis flowering time GWAS results using biological-informed priors (HapFM-anno). (a) The estimated effect sizes of different biological annotations for the Arabidopsis flowering time dataset. (b) The comparison of significant loci identified with and without incorporating biological annotations. (c) Manhattan plots HapFM-anno on Arabidopsis flowering time (FT10) dataset. The red dash line indicates the FDR 0.05 threshold. The size of the dots indicates the estimated effect size of the block.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## Discussion

GWAS has emerged as a critical approach to understanding the genetic architecture of complex traits and diseases especially in medical studies. Its utility in plant studies has been limited by a dearth of suitable genomic datasets. Yet, as the volume of plant genomic and phenotypic datasets increase, GWAS will begin to take on a more significant role as it does in human studies. SNP-based LMM and its variants are commonly used but often underpowered in plant GWAS studies due to limitations in the study designs and the high complexity nature of agronomic traits [3,40]. Conventional GWAS methods use LMM to identify significant SNPs by marginally testing one SNP at a time without considering LD between proximal SNPs.

There may be reasons why a conventional GWAS approaches may not be the most suitable model for plant GWAS. Plant GWAS generally have a small population size, a magnitude or two smaller than most human GWAS. In these circumstances, when an individual SNP has a large effect size, marginal regression can successfully identify it together with its in-LD SNPs and results in a significant peak in the Manhattan plot even in small GWAS populations. For instance, conventional GWAS methods have been used in small populations to map traits contributed by large-effect loci, such as qualitative resistance [41], plant architecture [4], metabolic pathways [5]. On the other hand, conventional GWAS methods often struggle to map traits contributed by numerous small-effect loci in populations of limited size. For example, significant SNPs identified by an LMM-based GWAS method, FarmCPU, only explained 15% of the phenotypic variation in a *Sclerotinia* resistance in soybean[42]. This result is consistent with our simulation results that GEMMA, a representative of conventional LMM-based GWAS method, that correctly identified large-effect loci in low-polygenicity traits while failing to identify small-effect loci in high polygenicity traits. One way of increasing mapping power is to increase sample size in GWAS. For example, in human height GWAS, 253,288 individuals were analyzed identifying 423 loci, with the majority loci contributing less than 1% of the total heritability [43]. Aggregating SNP effects is another way of increasing mapping power, such as SNP-set based method. This assumes that there may exist more

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

than one causal SNPs in the SNP-set. HapFM follows a similar strategy by projecting SNPs on haplotypes and then testing the effect sizes of haplotypes rather than individual SNPs. In addition, using haplotypes as variables also includes cis-interaction between SNPs, which is generally missing in SNP-based LMM models.

The second reason conventional GWAS models are underpowered is that a large number of SNPs cause multiple testing burdens in the marginal regression. As sequencing cost continues to decrease, however, genotyping a GWAS cohort by whole genome sequencing has become more affordable than ever before. When WGS datasets are used in plants, the high levels of genetic diversity of many plant species create datasets whereby millions of SNPs / INDELs can be identified in individuals, especially when including wild relatives [44]. This excessively large number of SNPs can affect the power of conventional SNP-based LMM methods because significance is tested on individual SNPs with overall significance calculated with cutoffs to control type I error. The overall significance cutoff will be more stringent as the number of SNPs increases in the analysis, significantly reducing the power of conventional SNP-based GWAS methods, such as GEMMA, GAPIT, and FarmCPU. A common solution to the multiple testing issue is to select a subset of representative SNPs for each LD block, also known as "tag SNPs", to reduce the number of tests in the analysis. This method assumes, however, that the causal SNPs are in LD with the tag SNPs [45] [46]. This can be problematic since the selection of the representative SNP is arbitrary involving choosing parameters for LD cutoff and physical distance. Moreover, information about other SNPs is lost with this method, such as the number of causal SNPs, LD structure between nearby SNPs. As discussed below, HapFM solves the multiple testing problem by combining SNPs into haplotypes, which greatly reduced the total number of variables in the model.

Another limitation of conventional GWAS methods is the interpretability of mapping results, including mapping interval and relevant biological information. Domestication and modern breeding result in large LD blocks in many crop genomes [47] and most conventional GWAS methods marginally test each SNP marker without considering the LD between nearby SNPs. Therefore, a bundle of proximal SNPs may pass the significance threshold simply due to strong regional LD, resulting in a large

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

significant peak in the Manhattan plot. This is especially problematic when the mapping interval of the locus is defined as the boundary where LD decays below a threshold ($r^2 < 0.1$). In a region with high LD, the mapping interval could span hundreds of genes and compounding the difficulty downstream experimental validation [3,8,23]. A common practice to increase mapping resolution in the high LD region in many plants is to generate a fine-mapping population to further reduce LD by introducing recombination into the region [48]. Nevertheless, developing a fine-mapping population is labor-intensive and at a high cost, which largely limits its application. Mapping resolution can also be improved by performing statistical fine-mapping in the region to identify a credible set of SNPs with a high probability containing the true causal SNPs. Statistical fine-mapping methods has been successfully used in human genetic studies to narrow down the list of causal SNPs [49 50]. One limitation of this method, however, is that it is locus-specific rather than genome-wide due to high computation intensity. Also, biological interpretation of the SNPs in the credible set may be ambiguous because they may not be obvious functional variants.

HapFM leverages the combination of genome-wide haplotype block fine-mapping with statistical fine-mapping to identify causal haplotype blocks. When possible, HapFM partitions large independent blocks into smaller and correlated blocks to further increase mapping resolution. LD information between small blocks is then used to identify the causal blocks. The causal block identified provides a reduced interval for the identification of functional variants. One limitation of this method, however, is that structural rearrangements, such as inversion, may result in the location of functional variants outside of the identified causal blocks.

Comparison with other GWAS methods in the simulation and real datasets showed that HapFM could greatly increase mapping resolution and achieve higher mapping power with complex traits. This indicates that HapFM may greatly improve current mapping efforts and perhaps serve as an alternative GWAS strategy in plant studies. Our results show that HapFM generated smaller mapping intervals than GEMMA, especially in regions of high LD in the simulation studies. HapFM consistently mapped traits to a smaller interval with fewer candidate genes than GEMMA. These results suggest that HapFM is capable of addressing the previously mentioned limitations found in many plant GWAS studies. In low

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

polygenitcity simulations, GEMMA showed higher mapping power than HapFM, suggesting GEMMA, or SNP-based LMM models in general, would provide a powerful method for mapping simple traits contributed by major effect loci. Therefore, the choice of the mapping algorithms may be determined by the genetic architecture of the traits. Other methods, such as GMMAT and BSLMM, consistently underperformed in both the simulation and actual plant datasets. Therefore, optimization of the models is necessary for better plant applications.

A similar haplotype-based method, FH-GWAS [15], has been developed which demonstrates an advantage of using haplotypes over SNP as variables by aggregating local epistatic effects. In our study, FH-GWAS and HapFM identified more significant loci than conventional SNP-based methods on the same Arabidopsis FT10 GWAS dataset (Supplemental Table 1). Overall, HapFM identified two more significant loci than FH-GWAS in the Arabidopsis FT10 GWAS dataset. The improved mapping power may be due to the following reasons. HapFM has benchmarked different block partitioning algorithms and showed the advantages of non-uniform LD-based partitioned using BigLD over uniform partitioning and PLINK partitions. HapFM goes further by performing haplotype clustering instead of using unique haplotypes, reducing the number of variables in the final model, and increasing the power of low-frequency haplotypes. Finally, HapFM uses the full model instead of marginal regressing haplotypes methods used in most haplotype-based GWAS methods, such as FH-GWAS and RAINBOW [14]. The full model doesn't need to estimate the kinship between individuals, and the output results from HapFM indicate causal signals. Last but not least, HapFM can use biological-informed priors for different genomic regions, which could further improve its mapping power.

One limitation of HapFM is its high computational time. This computational cost is determined by factors including the number of blocks in the genome, the sensitivity of haplotype clustering, and the number of MCMC iterations. HapFM uses the full model rather than marginal regression to infer the causality of each block. The more blocks partitioned, the more variables will be included in the fine-mapping model, which essentially increases resolution at the expense of computational intensity. Similarly, failing to cluster haplotypes will also increase the number of variables in the model. HapFM

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

uses MCMC for parameter inference, and the number of iterations for MCMC to reach convergence is random and highly variable. In addition, a large number of iterations is necessary to reduce the standard error of the estimates. These factors all contribute to the high computational time of HapFM.

Future improvements on HapFM include, but are not limited to, optimization in block partition and haplotype clustering algorithms and reducing computation time in the MCMC step. Moreover, as more and more plant species now have a pan-genome reference showing complex structural variations in different individuals [51], a pan-genome compatible trait mapping algorithm will be in high demand in the near future. The conventional SNP-based marginal regression models may struggle to be applied to the pan-genome reference because different reference genomes will output different sets of SNP genotypes as well as structural variations. HapFM has an advantage in pan-genome-based trait mapping because it uses haplotype as variables, defined by SNPs and structural variations. In addition, different reference genomes increase the accuracy and resolution of haplotype identification by providing extra information. The application of HapFM on pan-genome references is still under development.

In conclusion, we have developed a novel GWAS algorithm, HapFM, to address specific issues in plant studies. We demonstrated that HapFM showed advantages in shorter mapping intervals and higher mapping power than conventional GWAS methods in simulation and actual plant datasets. These results suggested that HapFM is a reliable alternative GWAS algorithm, and it supplements the current GWAS methods to facilitate the understanding of genetic architecture of traits.

## Material and Methods

*Genome-wide haplotype block partition*

HapFM first performs genome-wide block partitioning, outputting sets of non-overlapping SNPs using LD between SNPs as the partitioning metric. Previous studies have demonstrated that given the genotype data of a population, the linear reference genome can be divided into blocks with limited

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

haplotype diversity, also known as haplotype blocks [52]. HapFM utilizes a 2-step partitioning strategy to achieve high computation efficiency. The first step identifies large independent blocks which are defined as a proximal set of SNPs with minimum pairwise LD ($r^2$) that are larger than a pre-defined threshold ($r^2$=0.1 by default). A maximum distance threshold between SNP pairs is also set to avoid unrealistically large blocks caused by randomness. The second step in the partitioning process identifies sub-block structures within the large independent block by using existing block partition algorithms. The current version of HapFM has the choice of three block partition algorithms -- Uniform partition, PLINK [53] and BigLD [54]. Users can also input their own block partitions.

*Haplotype clustering*

After the block partition step, HapFM performs haplotype clustering on the unique haplotypes present in each haplotype block. In this clustering step, HapFM first enumerates all of the unique haplotypes in the block. When the number of unique haplotypes exceeds the user-defined threshold ($n$ = 10 by default), HapFM will perform haplotype clustering to reduce the number of variables in the mapping step. For a block containing $h$ unique haplotypes characterized by $s$ SNPs, HapFM uses the SNP indicator matrix ($h \times s$) as input for the clustering algorithms. HapFM currently has implemented four clustering methods: affinity propagation, X-means, local scaling (LS)-spectral clustering and K-nearest neighbor (KNN)-spectral clustering. Affinity propagation was implemented using sklearn.cluster.AffinityPropagation function from the scikit-learn package (0.23.2). X-means was implemented using the X-Means function from the Pyclustering library [55]. LS-Spectral clustering and KNN-Spectral clustering were implemented using in-house python scripts.

*Genome-Wide Haplotype Fine Mapping Model*

The genome-wide haplotype fine mapping model follows a linear mixed model (LMM) and a hierarchical Bayes inference framework:

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$y = \mathbf{C}\alpha + \mathbf{H}\beta + \epsilon,$$

where $y$ is a length $n$ vector of phenotypic values; $\mathbf{C}$ is an $n \times c$ matrix of covariates, $\alpha$ is a length $c$ vector containing the fixed effects of covariates; $\mathbf{H}$ is an $n \times m$ design matrix indicating the counts of haplotype (clusters); $\beta$ is a length $m$ vector of random effects of haplotype (clusters); $\epsilon$ is a length $n$ vector of random residual effects. The prior distribution for effect size $\beta$ is shown as below:

$$\beta \sim (1-\pi)N(0,\delta_0^2) + \pi N(0,\delta_1^2),$$

$$\beta_i \mid \gamma_i \sim \begin{cases} N(0,\delta_0^2) & \text{if } \gamma_i = 0 \\ N(0,\delta_1^2) & \text{if } \gamma_i = 1 \end{cases},$$

$$\gamma_i \sim \text{Bernoulli}(\pi),$$

$$\delta_1^{-2} \sim \text{Gamma(a, b)},$$

$$\beta_{PIP} = \text{E}(\gamma \mid \text{y}, \mathbf{H})$$

As shown in the model, the haplotype effect sizes follow a mixture of normal density with mean 0 and variance $\sigma_1^2$ and a normal density with variance $\sigma_0^2$ pre-specified close to 0. The latent variable $\gamma$ encodes the components whose corresponding effect size come from $N(0,\sigma_1^2)$ . The inference was performed using an in-house Gibbs sampler, and the posterior inclusion probability (PIP) of each $\beta$ indicates the inferred probability of the haplotype block being causal.

The parameter $\pi$ suggests the prior probability of causality for each haplotype block. If annotation is not provided, the model assumes every haplotype block has the same prior probability for causality. If biological annotations are provided, the causal probability of each haplotype block will be inferred by fitting it into the following Probit model:

$$\Phi^{-1}\left[P(\gamma_i = 1)\right] = \mathbf{A}^T\theta,$$

where $\Phi^{-1}$is the inverse of cumulative distribution function of a standard normal distribution, $\mathbf{A}$ is the matrix containing the annotation features, and $\theta$ is the vector of effect size corresponding to each biological annotation. The inference of $\theta$ follows the data augmentation method from [56].

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

*Simulation analyses*

Simulation datasets were generated to compare different block partition and haplotype clustering algorithms implemented in the HapFM framework and to benchmark the mapping performance of HapFM against conventional GWAS methods.

In genotype simulation, populations with 500 individuals were simulated to contain 100 large independent blocks in the genome. In each large independent block, the number and the size of sub-blocks, $s$, was sampled from the Uniform (1, 10) distribution and Uniform (10, 100) distribution, respectively. The number of haplotype clusters, $h_c$, in each sub-block was randomly sampled from a Uniform (2, 4) distribution. Haplotype diversity, $d$, is a parameter to simulated different diversity of the simulated population. The total number of unique haplotypes, $h$, was calculated as $h_c \times d$. Random mutations were then introduced to haplotype clusters to generate unique haplotypes. The unique haplotype matrix $Z^{h \times s}$ encompassed the SNP features of all the haplotypes in the block. The haplotype frequencies, $f_h$, were calculated by solving the linear equation:

$$f_s = Z f_h$$

whereby the $f_s$ is a vector of the minor allele frequencies in the block randomly sampled from a Uniform (0.05, 0.95) distribution. The haplotypes were then sampled from a Multinomial (2, $f_h$) to generate the genotype of the block for each individual.

The phenotype of the population was simulated using the following equation:

$$y = \boldsymbol{C}\alpha + \boldsymbol{X}\eta + \epsilon,$$

whereby the coefficients $\alpha$ were sampled from a Uniform (-1, 1) distribution, and the entries in the covariate matrix $C$ were sampled from a Uniform (-5, 5) distribution. $\boldsymbol{X}$ represents the simulated SNP genotype matrix. $\eta$ represents the SNP effect sizes which was simulated in a hierarchical manner: casual blocks and casual SNPs in the block. At the block level, the probability, $\pi_B$, of a block containing true causal SNPs was simulated at 0.005 and 0.05. and the block effect size $\eta_{B_j}$ were simulated ranging from

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

0.5 to 3. Under each true causal block, four types of architectures of true causal SNPs ($\lambda_i = 1$) were simulated (Figure 1a):

(1) Architecture No.1: one large effect causal SNP;

(2) Architecture No.2: Five or six small effect causal SNPs randomly assigned to haplotypes;

(3) Architecture No.3: two moderate effect causal SNPs assigned to different haplotypes;

(4) Architecture No.4: mixture of large and small effect causal SNPs randomly assigned to haplotypes;

For each architecture, SNP-level effect size, $\eta_i$, was assigned to each individual causal SNP based on the equation $\beta_{B_j} = \sum_{SNP_i \in B_j} \beta_i \mathbf{I}(\lambda_i = 1)$, where $\mathbf{I}$ is the indicator function. The effect sizes of non-causal SNPs were randomly sampled from the Normal $(0, 0.0001)$ distribution.

*Processing of real datasets*

In real data analyses, five existing datasets were used to demonstrate the performance of HapFM on various types of genetic architectures and LD structures, and benchmark it with other GWAS method. These datasets were an Arabidopsis flowering time dataset (FT10) [57], tomato metabolite [58], rice yield [59], maize height [60] and a cassava HCN content [61]. The Arabidopsis flowering time GWAS dataset included genotype information from two previously published datasets: Arabidopsis Regmap [62] and 1001 Arabidopsis genome [63]. In the 1001 Arabidopsis genotype dataset, non-biallelic SNPs and SNPs with missing percentage greater than 20% were filtered out giving a total of 8,231,757 remaining SNPs. In the Regmap genotype dataset, SNPs that are not in LD ($R2 < 0.1$) with nearby 20 SNPs we filtered out leaving 202,339 remaining SNPs, 170,977 of which were also included in the filtered 1001 Arabidopsis genotype dataset. The overlapping SNPs were used as the reference panel for imputation using Beagle 4.1 [64] to impute missing data and phased genotypes by following a 2-step imputation procedure [44]. After imputation and phasing, SNPs with a minor allele frequency (MAF) $< 0.05$ and those that were not in LD with nearby 20 SNPs were removed resulting in a 1,013,248 final SNPs dataset. Next, genome-wide LD

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

pruning was performed on the filtered genotypes using PLINK with parameter set as --indep-pairwise 1000 100 0.1 [65]. Finally, principal component analysis (PCA) was performed on LD-pruned SNPs and the first five PCs were used as covariates to adjust for population structure.

The tomato fruit metabolic GWAS dataset was downloaded from published data [58]. The genotype data of the 441 tomato accessions were processed according to Wu et al. published workflow [44]. A total of 3,281,705 SNPs were kept after filtering out SNPs with MAF < 0.05 and SNPs that were not in LD ($r^2$ < 0.1) with nearby 20 SNPs. Genome-wide LD pruning was then performed using PLINK with parameter set as --indep-pairwise 1000 100 0.1 and remained 7,747 LD-pruned SNPs. The first two PCs were used as covariates to adjust for population structure. The concentration of SlFM0969 metabolite, Apigenin 7-O-glucoside, was used for the phenotype in the analysis.

The genotype and yield phenotype datasets of 295 rice individuals were downloaded from Rice Variation Map ( http://ricevarmap.ncpgr.cn/ ) [66]. Beagle 4.1 was used to impute missing data and to phase genotypes. A total of 1,017,380 SNPs were used for GWAS analysis after removing SNPs with MAF < 0.05 and SNPs that were not in LD ($r^2$ < 0.1) with nearby 20 SNPs. Genome-wide LD pruning was then performed on the filtered rice genotypes using PLINK with parameter set as --indep-pairwise 1000 100 0.1 and remained 12367 LD-pruned SNPs. PCA was performed on LD-pruned SNPs and the first two PCs were used as covariates to adjust for population structure.

The genotype information and HCN content of 1239 cassava accessions were obtained from a published dataset [61]. A total of 16596 SNPs were kept for GWAS analysis after filtering out SNPs with MAF < 0.05 and SNPs that were not in LD ($r^2$ < 0.1) with nearby 20 SNPs. Genome-wide LD pruning was then performed using PLINK with parameter set as --indep-pairwise 1000 100 0.1 and remained 826 LD-pruned SNPs. PCA was performed on LD-pruned SNPs and the first 10 PCs were used as covariates to adjust for population structure.

The maize HapMapV3.2.1 genotypes and 263 plant height phenotypes were downloaded from Panzea ( https://www.panzea.org/ ). Beagle 4.1 was used to impute missing data and to phase genotypes. A total of 23,093,292 SNPs were used for GWAS analysis after removing SNPs with MAF < 0.05 and

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SNPs that were not in LD ($r^2 < 0.1$) with nearby 20 SNPs. Genome-wide LD pruning was then performed on the filtered rice genotypes using PLINK with parameter set as --indep-pairwise 1000 100 0.1 and remained 148,961 LD-pruned SNPs. PCA was performed on LD-pruned SNPs and the first three PCs were used as covariates to adjust for population structure.

*Benchmark different GWAS methods on simulated and real datasets*

In both simulation and real data analyses, HapFM was compared with three GWAS methods: traditional LMM-based univariate association mapping GEMMA v0.98.1 [67], Bayesian Sparse LMM BSLMM v0.98.1 [68], and SNP-set based association method SMMAT v1.3.1 [13]. The kinship matrix, if needed, was calculated by GEMMA with parameter -gk 1. To fit a univariate linear mixed model in GEMMA, corresponding covariates were used with default settings for the other parameters. To fit the BSLMM model, the -bslmm 1 option was used with default settings for the other parameters. No covariate was included in the BSLMM model. To fit the SMMAT model, SNP sets based on the haplotype blocks identified by HapFM used including the corresponding covariates and default settings all parameters.

In both simulation and real data analyses, the mapping power and mapping interval of different GWAS methods was compared with FDR set at < 0.05. HapFM and GMMAT identify significant haplotype blocks whereas BSLMM and GEMMA identify significant SNPs. Therefore, the FDR values for BSLMM and GEMMA results need to be adjusted to achieve a fair comparison. To do this, the most significant SNP in each HapFM block partition was selected as the representative SNP and the adjusted FDR values were calculated using the formular [69]:

$$\frac{|S|q}{M},$$

whereby $|S|$ represents the number of representative SNPs, $q$ represents the desired FDR level, and M represents the total number of SNPs. The mapping intervals of significant loci (FDR < 0.05) of each GWAS method were then calculated. The mapping intervals of HapFM and GMMAT were the length of

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

their corresponding blocks. The mapping interval of GEMMA and PLINK were calculated by clumping SNPs based on their pairwise LD using PLINK with the parameter set as --clump-r2 0.2. In addition, the mapping accuracy in the simulated study was calculated as the percentage of true positive blocks (FDR < 0.05) from each GWAS method. The blocks contained significant SNPs identified by GEMMA and BSLMM were used to calculate the accuracy of GEMMA and BSLMM, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## References


1. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).
2. Farashi, S., Kryza, T., Clements, J. & Batra, J. Post-GWAS in prostate cancer: from genetic association to biological contribution. *Nat Rev Cancer* **19**, 46-59 (2019).
3. Cortes, L.T., Zhang, Z.W. & Yu, J.M. Status and prospects of genome-wide association studies in plants. *Plant Genome* **14**(2021).
4. Yano, K. *et al.* GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. *Proc Natl Acad Sci U S A* **116**, 21262-21267 (2019).
5. Tieman, D. *et al.* A chemical genetic roadmap to improved tomato flavor. *Science* **355**, 391-394 (2017).
6. Li, N. *et al.* Natural variation in ZmFBL41 confers banded leaf and sheath blight resistance in maize. *Nat Genet* **51**, 1540-1548 (2019).
7. Huang, X.H. & Han, B. Natural Variations and Genome-Wide Association Studies in Crop Plants. *Annual Review of Plant Biology, Vol 65* **65**, 531-551 (2014).
8. Zhou, X. & Huang, X. Genome-wide Association Studies in Rice: How to Solve the Low Power Problems? *Mol Plant* **12**, 10-12 (2019).
9. Xiao, Y., Liu, H., Wu, L., Warburton, M. & Yan, J. Genome-wide Association Studies in Maize: Praise and Stargaze. *Mol Plant* **10**, 359-374 (2017).
10. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* **44**, 825-30 (2012).
11. Wang, Q., Tian, F., Pan, Y., Buckler, E.S. & Zhang, Z. A SUPER powerful method for genome wide association study. *PLoS One* **9**, e107684 (2014).
12. Liu, X., Huang, M., Fan, B., Buckler, E.S. & Zhang, Z. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLoS Genet* **12**, e1005767 (2016).
13. Chen, H. *et al.* Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. *Am J Hum Genet* **104**, 260-274 (2019).
14. Hamazaki, K. & Iwata, H. RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. *PLoS Comput Biol* **16**, e1007663 (2020).
15. Liu, F., Schmidt, R.H., Reif, J.C. & Jiang, Y. Selecting Closely-Linked SNPs Based on Local Epistatic Effects for Haplotype Construction Improves Power of Association Mapping. *G3 (Bethesda)* **9**, 4115-4126 (2019).
16. Badouin, H. *et al.* The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148-152 (2017).
17. Zhou, Z. *et al.* Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol* **33**, 408-14 (2015).
18. Lin, T. *et al.* Genomic analyses provide insights into the history of tomato breeding. *Nat Genet* **46**, 1220-6 (2014).
19. Ingvarsson, P.K. & Street, N.R. Association genetics of complex traits in plants. *New Phytol* **189**, 909-922 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

20. Li, B. *et al.* Identification and fine mapping of a major locus controlling branching in Brassica napus. *Theor Appl Genet* **133**, 771-783 (2020).
21. Wang, B. *et al.* Identification and Fine-Mapping of a Major Maize Leaf Width QTL in a Re-sequenced Large Recombinant Inbred Lines Population. *Front Plant Sci* **9**, 101 (2018).
22. Wang, Y. *et al.* Fine mapping of a major locus controlling plant height using a high-density single-nucleotide polymorphism map in Brassica napus. *Theor Appl Genet* **129**, 1479-91 (2016).
23. Schaid, D.J., Chen, W. & Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491-504 (2018).
24. Goktay, M., Fulgione, A. & Hancock, A.M. A New Catalog of Structural Variants in 1,301 A. thaliana Lines from Africa, Eurasia, and North America Reveals a Signature of Balancing Selection at Defense Response Genes. *Mol Biol Evol* **38**, 1498-1511 (2021).
25. Fuentes, R.R. *et al.* Structural variants in 3000 rice genomes. *Genome Res* **29**, 870-880 (2019).
26. Alonge, M. *et al.* Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* **182**, 145-161 e23 (2020).
27. Anderson, J.E. *et al.* A roadmap for functional structural variants in the soybean genome. *G3 (Bethesda)* **4**, 1307-18 (2014).
28. Yang, N. *et al.* Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat Genet* **51**, 1052-1059 (2019).
29. Kawakatsu, T. *et al.* Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell* **166**, 492-505 (2016).
30. Frey, B.J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972-976 (2007).
31. Pelleg, D. & Moore, A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *In Proceedings of the 17th International Conf. on Machine Learning*, 727-734 (2000).
32. Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing* **17**, 395-416 (2007).
33. Wang, Y. *et al.* Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in Arabidopsis. *Plant Physiol* **148**, 1201-11 (2008).
34. Jakoby, M.J. *et al.* Transcriptional profiling of mature Arabidopsis trichomes reveals that NOECK encodes the MIXTA-like transcriptional regulator MYB106. *Plant Physiol* **148**, 1583-602 (2008).
35. Pagnussat, G.C. *et al.* Genetic and molecular identification of genes required for female gametophyte development and function in Arabidopsis. *Development* **132**, 603-14 (2005).
36. Klepikova, A.V., Kasianov, A.S., Gerasimov, E.S., Logacheva, M.D. & Penin, A.A. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J* **88**, 1058-1070 (2016).
37. Wang, X. *et al.* Overexpression of PGA37/MYB118 and MYB115 promotes vegetative-to-embryonic transition in Arabidopsis. *Cell Res* **19**, 224-35 (2009).
38. Narsai, R., Law, S.R., Carrie, C., Xu, L. & Whelan, J. In-depth temporal transcriptome profiling reveals a crucial developmental switch with roles for RNA processing and

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

organelle metabolism that are essential for germination in Arabidopsis. *Plant Physiol* **157**, 1342-62 (2011).

39. Shinohara, N., Ohbayashi, I. & Sugiyama, M. Involvement of rRNA biosynthesis in the regulation of CUC1 gene expression and pre-meristematic cell mound formation during shoot regeneration. *Front Plant Sci* **5**, 159 (2014).
40. Korte, A. & Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9**, 29 (2013).
41. Tran, D.T., Steketee, C.J., Boehm, J.D., Jr., Noe, J. & Li, Z. Genome-Wide Association Analysis Pinpoints Additional Major Genomic Regions Conferring Resistance to Soybean Cyst Nematode (Heterodera glycines Ichinohe). *Front Plant Sci* **10**, 401 (2019).
42. Wei, W. *et al.* Genome-wide association mapping of resistance to a Brazilian isolate of Sclerotinia sclerotiorum in soybean genotypes mostly from Brazil. *BMC Genomics* **18**, 849 (2017).
43. Chan, Y. *et al.* Genome-wide Analysis of Body Proportion Classifies Height-Associated Variants by Mechanism of Action and Implicates Genes Important for Skeletal Development. *Am J Hum Genet* **96**, 695-708 (2015).
44. Wu, X., Heffelfinger, C., Zhao, H. & Dellaporta, S.L. Benchmarking variant identification tools for plant diversity discovery. *BMC Genomics* **20**, 701 (2019).
45. Wang, S., He, S., Yuan, F. & Zhu, X. Tagging SNP-set selection with maximum information based on linkage disequilibrium structure in genome-wide association studies. *Bioinformatics* **33**, 2078-2081 (2017).
46. Ding, K. & Kullo, I.J. Methods for the selection of tagging SNPs: a comparison of tagging efficiency and performance. *Eur J Hum Genet* **15**, 228-36 (2007).
47. Doebley, J.F., Gaut, B.S. & Smith, B.D. The molecular genetics of crop domestication. *Cell* **127**, 1309-21 (2006).
48. Jaganathan, D., Bohra, A., Thudi, M. & Varshney, R.K. Fine mapping and gene cloning in the post-NGS era: advances and prospects. *Theoretical and Applied Genetics* **133**, 1791-1810 (2020).
49. Westra, H.J. *et al.* Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. *Nat Genet* **50**, 1366-1374 (2018).
50. Ferreiro-Iglesias, A. *et al.* Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. *Nat Commun* **9**, 3927 (2018).
51. Lei, L. *et al.* Plant Pan-Genomics Comes of Age. *Annu Rev Plant Biol* **72**, 411-435 (2021).
52. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).
53. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
54. Kim, S.A., Cho, C.S., Kim, S.R., Bull, S.B. & Yoo, Y.J. A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics* **34**, 388-397 (2018).
55. Novikov, A.V. PyClustering: Data Mining Library. *Journal of Open Source Software* **4**, 1230 (2019).
56. Albert, J.H. & Chib, S. Bayesian-Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* **88**, 669-679 (1993).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.28.466332; this version posted October 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

57. Seren, U. *et al.* AraPheno: a public database for Arabidopsis thaliana phenotypes. *Nucleic Acids Res* **45**, D1054-D1059 (2017).
58. Zhu, G. *et al.* Rewiring of the Fruit Metabolome in Tomato Breeding. *Cell* **172**, 249-261 e12 (2018).
59. Xie, W. *et al.* Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc Natl Acad Sci U S A* **112**, E5411-9 (2015).
60. Peiffer, J.A. *et al.* The genetic architecture of maize height. *Genetics* **196**, 1337-56 (2014).
61. Ogbonna, A.C. *et al.* Large-scale genome-wide association study, using historical data, identifies conserved genetic architecture of cyanogenic glucoside content in cassava (Manihot esculenta Crantz) root. *Plant J* **105**, 754-770 (2021).
62. Horton, M.W. *et al.* Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nat Genet* **44**, 212-6 (2012).
63. Genomes Consortium. Electronic address, m.n.g.o.a.a. & Genomes, C. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481-491 (2016).
64. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).
65. Borile, C., Labarre, M., Franz, S., Sola, C. & Refregier, G. Using affinity propagation for identifying subspecies among clonal organisms: lessons from M. tuberculosis. *BMC Bioinformatics* **12**, 224 (2011).
66. Zhao, H. *et al.* RiceVarMap: a comprehensive database of rice genomic variations. *Nucleic Acids Res* **43**, D1018-22 (2015).
67. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821-4 (2012).
68. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet* **9**, e1003264 (2013).
69. Brzyski, D. *et al.* Controlling the Rate of GWAS False Discoveries. *Genetics* **205**, 61-75 (2017).