Spring 2022

Uncovering Hidden Diversity in Plants

Xing Wu
Yale University Graduate School of Arts and Sciences, wuxingtom@gmail.com

Abstract

Uncovering Hidden Diversity In Plants

Xing Wu

2022

One of the greatest challenges to human civilization in the 21st century will be to provide global food security to a growing population while reducing the environmental footprint of agriculture. Despite increasing demand, the fundamental issue of limited genetic diversity in domesticated crops provides windows of opportunity for emerging pandemics and the insufficient ability of modern crops to respond to a changing global environment. The wild relatives of crop plants, with large reservoirs of untapped genetic diversity, offer great potential to improve the resilience of elite cultivars. Utilizing this diversity requires advanced technologies to comprehensively identify genetic diversity and understand the genetic architecture of beneficial traits. The primary focus of the dissertation is developing computational tools to facilitate variant discovery and trait mapping for plant genomics.

In Chapter 1, I benchmarked the performance of variant discovery algorithms based on simulated and diverse plant datasets. The comparison of sequence aligners found that BWA-MEM consistently aligned the most plant reads with high accuracy, whereas Bowtie2 had a slightly higher overall accuracy. Variant callers, such as GATK HaplotypCaller and SAMtools mpileup, were shown to significantly differ in their ability to minimize the frequency of false negatives and maximize the discovery of true positives. A cross-reference experiment of *Solanum lycopersicum* and *Solanum pennellii* reference genomes revealed significant limitations of using a single reference genome for variant discovery. Next, I demonstrated that a machine-learning-based variant filtering strategy outperformed the traditional hard-cutoff filtering strategy, resulting in a significantly higher number of true positive and fewer false-positive variants. Finally, I developed a 2-step imputation method resulted in up to 60% higher accuracy than direct LD-based imputation methods.

In Chapter 2, I focused on developing a trait mapping algorithm tailored for plants considering the high levels of diversity found in plant datasets. This novel trait mapping framework, HapFM, had the ability to incorporate biological priors into the mapping model to identify causal haplotypes for traits of interest. Compared to conventional GWAS analyses, the haplotype-based approach significantly reduced the number of variables while aggregating small effect SNPs to increase mapping power. HapFM could account for LD between haplotype segments to infer the causal haplotypes directly. Furthermore, HapFM could systemically incorporate biological priors into the probability function during the mapping process resulting in greater mapping resolution. Overall, HapFM achieves a balance between powerfulness, interpretability, and verifiability.

In Chapter 3, I developed a computational algorithm to select a pan-genome cohort to maximize the haplotype representativeness of the cohort. Increasing evidence suggest that a single reference genome is often inadequate for plant diversity studies due to extensive sequence and structural rearrangements found in many plant genomes. HapPS was developed to utilize local haplotype information to select the reference cohort. There are three steps in HapPS, including genome-wide block partition, representative haplotype identification, and genetic algorithm for reference cohort selection. The comparison of HapPS with global-distance-based selection showed that HapPS resulted in significantly higher block coverage in the highly diverse genic regions. The GO-term enrichment analysis of the highly diverse genic region identified by HapPS showed enrichment for genes involved in defense pathways and abiotic stress, which might identify genomic regions involved in local adaptation. In summary, HapPS provides a systemic and objective solution to pan-genome cohort selection.

Uncovering Hidden Diversity In Plants

A Dissertation

Presented to the Faculty of the Graduate School

Of

Yale University

In Candidacy for the Degree of

Doctor of Philosophy

By

Xing Wu

Dissertation Director: Dr. Stephen Dellaporta and Dr. Hongyu Zhao

May 2022

Copyright page

Table of Contents

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Stephen Dellaporta and Prof. Hongyu Zhao, for their continuous support of my Ph.D. study. Specifically, I want to thank Prof. Stephen Dellaporta for his patience, motivation, inspiration, and caring throughout my PhD research and personal life. I want to thank Prof. Hongyu Zhao for his immense knowledge and providing a collaborative environment to help with my Ph.D. study. Their guidance helped me in all the time of my research, preparing for presentations, and writing of this thesis, and made my Ph.D. enjoyable and productive.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Joshua Gendron, Prof. Zuoheng Wang, and Prof. John Lafferty for their insightful comments, allowing me to widen my research from various perspectives.

I want to thank my parents, Dr. Junde Wu and Chunxiang Liu, for their love and support throughout my life. Finally, I would like to thank my fiancée, Dr. Wei Wei and my friends at Yale University. Without them, I would not have made this far.

Statement of Collaborations and Contributions

This statement acknowledges people I have collaborated with on my research and their contributions.

In Chapter 1, I collaborated with Christopher Heffelfinger on the cross-reference comparison of the variant discovery pipeline. He helped me analyze the syntenic regions between two tomato reference genomes using the SNP data.

In Chapter 2, I collaborated with Dr. Wei Jiang, Dr. Chris Fragoso, Jing Huang, Geyu Zhou on haplotype-based trait fine mapping. Dr. Wei Jiang helped me formulate the statistical model for mapping, gave me suggestions on the block partition and haplotype cluster algorithms, and discussed the results. Dr. Chris Fragoso helped me with the haplotype clustering algorithm and preparing the rice dataset for my benchmarking analysis. Jing Huang and I worked on simulating genotypes and haplotypes. Geyu Zhou helped me with the independent block partition algorithm.

In Chapter 3, I collaborated with Dr. Wei Jiang. Dr. Wei Jiang helped me with the Genetic Algorithm and discussed the results.

Introduction

One of the greatest challenges to human civilization in the 21st century is providing global food security to a growing population while mitigating crop losses and agriculture's environmental footprint. As the global population grows, there is an increased pressure on the agricultural industry. Environmental factors, such as plant pathogens and climate changes, pose a significant threat to global agricultural production. Ug99 wheat rust, for example, emerged in Uganda in 1999, then spread into East Africa and Central Asia, and caused nearly 100% of the wheat yield loss yearly (Figueroa et al., 2018). With no genetic resistance found in domesticated germplasms, the prospects of worldwide losses loomed. Fortunately, natural resistance to Ug99 was identified in wild wheat relatives, alleviating the pathogen's global impact (Saintenac et al., 2013). Solutions to other pandemics have not yet been realized. Citrus greening disease, for instance, is an incurable disease regarded as the most devastating pandemic to the worldwide production of citrus. In Florida alone, one of the country's largest orange-producing states, crop loss has reached 74% since the disease was first identified (Singerman and Rogers, 2020). Commercial banana production is also seriously threatened by Panama disease caused by the soil-borne fungus *Fusarium oxysporum*. Currently, the exclusion of the pathogen is the primary measure to protect banana production (Ordonez et al., 2015). These are just a few examples whereby natural resistance to the disease has not been identified in commercial gene pool, emphasizing the importance of identifying new sources of diversity for global food security.

Crop production is also vulnerable to environmental challenges due to climate change, such as availability of water supply, temperature change, and the occurrence of extreme weather events. For each degree of global temperature increase, wheat yield is predicted to decrease by 4-6%. Approximately 56% of the worldwide maize-producing areas may also experience yield declines due to climate change by the end of the century (Anderson et al., 2020) (Anderson et al., 2020) (Zhao et al., 2017). These crop losses are compounded by the emergence of plant diseases associated with changing environmental conditions emphasizing the increased importance of resilient crop varieties.

Loss of genetic diversity is one of the underlying reasons why modern crops are vulnerable to emerging disease and environmental stress. Throughout the history of crop domestication, wild species were bred into modern varieties that represented a series of genetic bottlenecks that vastly reduced the genetic diversity base found in modern crop varieties (Gross and Olsen, 2010; Heslop-Harrison and Schwarzacher, 2012; Smith et al., 2015). Domesticated traits, such as yield, flowering time, and plant architecture were selected, leaving behind a reservoir of untapped genetic diversity in the wild relatives. This untapped diversity may contain a source of new diversity needed to provide resistance to emerging pathogens and environmental threats (Kofsky et al., 2018). For example, the wild soybean species *Glycine soja* has resistance to a severe pathogen, Asian Soybean Rust, and abiotic resistance to adapt to harsh environments (Kofsky et al., 2018). The infusion of this natural genetic resistance into modern breeding populations represent a source of untapped diversity that may offer protection from emerging threats to food security in the 21st-century and beyond.

Numbers of studies have shown landraces and wild relatives of our crops represent a source of genetic diversity that can be utilized for crop improvement. The resequencing of 1,143 indica rice accessions identified a new allele in OsNramp5 that can mitigate grain cadmium levels through hybrid breeding (Lv et al., 2020b). The resequencing study of 610 tomato accession, including 42 accessions of wild species and 568 accessions from the red-fruited clade, identified over 26 million SNPs in total. This multi-omics tomato flavor study provided a framework on how introgression from wild relatives could improve flavor and desirable metabolic changes (Zhu et al., 2018). In another study, resequencing 302 wild and cultivated soybean accessions uncovered 13 previously uncharacterized loci for agronomic traits including oil content, plant height and pubescence form (Zhou et al., 2015). These large resequencing studies further underscore the importance of exploring the reservoir of useful genetic diversity found in wild accessions.

The first step of utilizing these resources is to catalog the diversity found in the gene pool accurately and comprehensively. Current genomic resources, especially wild species, are not systemically exploited. Lack of resources represent a significant barrier for plant breeders and molecular biologists.

The two main obstacles are 1) that existing variant identification workflows are not optimized for the vast diversity found in plant species but rather for the much less diverse human genome; and 2) a single reference genome is an incomplete tool to represent the extent of genetic diversity of a plant species (Computational Pan-Genomics, 2018; Paten et al., 2017).

The first step in identifying diversity is a genomic workflow for variant discovery that involves genome sequencing, read mapping, raw variant identification, variant filtering, and missing data imputation. Many of the programs available for this pipeline were established and optimized for human studies, but few were specifically tailored for plant datasets (Krishnan et al., 2021; Kumaran et al., 2019; Liang et al., 2019). As compared to the human genome, plant genomes are often of lower quality, much more diverse, contain genetic redundancy due to polyploidy, and contain a higher fraction of highly repetitive regions as compared to human genomes. For example, the repetitive sequence content in the soybean genome and maize genome is around 57% and 80%, respectively (Haberer et al., 2020; Schmutz et al., 2010). In contrast, the repetitive sequence content in the human genome is around 45% (Venter et al., 2001). These differences contribute to the lower performance of variant discovery algorithms when used on plant datasets compared to humans. Few benchmarking studies on variant discovery pipelines focused on closely related individuals or populations with small diversity (Highnam et al., 2015; Krishnan et al., 2021; Kumaran et al., 2019; Liang et al., 2019; Schilbert et al., 2020). There is an immediate need to understand the performance of programs in the variant discovery pipeline on distantly related individuals and diverse populations.

In Chapter 1, I performed a benchmarking study on various programs in the variant discovery pipeline to address the question: how do different bioinformatic programs handle diversity. The benchmark included programs in the sequence read alignment, variant calling, variant filtering, and imputation steps using both simulated and real plant genomic datasets. A comparison of sequence aligners showed significant differences in alignment accuracy and alignment percentage among different aligners in distantly related individuals. The levels of diversity, sequence coverage, and genome complexity significantly affected the false-negative and true-positive rates of different variant calling programs. In the diversity population situations, the machine-learning-based variant filtering strategy outperformed the traditional hard-cutoff strategy resulting in a higher number of true positive variants and fewer false-positive variants. Finally, the 2-step imputation method, which utilized a set of high-confidence SNPs as the reference panel, showed up to 60% higher accuracy than direct LD-based imputation when analyzing wild species. In conclusion, this study serves as a piece of important guiding information for plant biologists utilizing next-generation sequencing data for diversity characterization and crop improvement.

Diversity without trait association cannot solve specific food security challenges associated with emerging pandemics and climate mitigation. Centuries of breeding turning weeds into productive crops is a selective process by which undesirable variation is eliminated while desirable agronomic traits are fixed (Gross and Olsen, 2010; Heslop-Harrison and Schwarzacher, 2012). Nevertheless, as new diseases appear and our planet's climate changes, this selective process may not have proactively retained useful variation relevant to these emerging challenges. Therefore, associating novel variation with traits of interest is critical in utilizing genetic diversity for crop improvement.

This critical association process of trait identification represents the starting point for the functional characterization of diversity. Part of this process is to define the genetic architecture of a trait of interest. For instance, the trait gene discovery that defines the genetic architecture of agronomically useful genetic diversity may involve simple Mendelian traits or more complex quantitative traits involving several pathways (Holland, 2007). In the field of genome editing, trait architecture is a critical process in identifying editing targets for crop improvement (Chen et al., 2019c). Gene editing is only a practical solution to the extent that the genetic architecture of a desirable trait is well defined and understood. If the edited genes do not contribute to the trait's heritability, varietal improvement will not occur. Crop improvement by gene editing promises to incorporate beneficial genetic diversity in a single generation while avoiding the downfalls of linkage drag that may take years to remove by recurrent selection.

Accordingly, a significant portion of plant research is dedicated to characterizing the segregating trait and its underlying genetic architecture. In addition to being critical for gene discovery, trait mapping is a foundational process for varietal improvement across the entire agricultural industry. As one would expect, plant researchers have applied considerable resources towards developing genomic resources and datasets surrounding the mapping and introgression of genetic diversity present in crop species, with a particular focus on the allelic diversity of trait genes. For example, a genome-wide association study (GWAS) of rice blast disease identified resistance loci and revealed a complex relationship between disease resistance and yield-related components (Wang et al., 2014a; Wang et al., 2015). The GWAS of seed protein and oil content in soybean identified 25 SNPs in 17 and 13 different genomic regions associated with seed protein and oil content, respectively (Hwang et al., 2014). Different from the diverse GWAS population, nested association mapping population (NAM) was used to study the genetic architecture of maize flowering time, and found numerous small-effect loci contributing to the phenotypic variation (Buckler et al., 2009). In many of these studies, especially those involving intricate population design or complex traits, there is emerging evidence that key genes are not being identified due to either insufficient mapping power or the lack of a suitable reference genome (Brachi et al., 2011; Cortes et al., 2021; Huang and Han, 2014b; Ingvarsson and Street, 2011; Visscher et al., 2017; Xiao et al., 2017; Zhou and Huang, 2019). For instance, plant QTL mapping often relies on a relatively small mapping population presenting challenges to sufficient mapping power and resolution to identify the causal genes. Developing and maintaining a mapping population for each trait of interest can be time-consuming and labor-intensive. On the other hand, when sufficient populations are available, GWAS is a commonly used method to define trait architecture as it provides higher power than QTL mapping, especially to polygenic traits. Nevertheless, GWAS was established and optimized using a human genomic dataset often encompassing tens of thousands of cohorts and extensive genomic datasets and has moved beyond just association in human genomic studies (Gallagher and Chen-Plotkin, 2018; Visscher et al., 2017). The small number of plant cohorts, limited data availability, high diversity, and large linkage-disequilibrium (LD) regions considerably reduces plant GWAS studies' efficacy. Therefore, it is necessary to develop a trait mapping algorithm suitable for the plant dataset, given its unique data structures.

In Chapter 2, I develop a plant-focused mapping algorithm that delivers high mapping power and resolution designed to broadly accelerate crop improvement. The algorithm, HapFM, uses a haplotype-based fine-mapping framework to address the issues in conventional plant GWAS studies. HapFM framework contains 4 steps: genome-wide block partition, unique haplotype counting, haplotype clustering, and genome-wide statistical fine-mapping with biological-informed priors. In the simulated datasets, I compared to the performance of different block partition and haplotype clustering algorithms. The results showed that BigLD, an LD-informed partitioning method, and X-means, a haplotype clustering algorithm, resulted in the highest mapping power and computational efficiency. The proof-of-concept study showed biological-informed priors could further increase the mapping power of HapFM using the Arabidopsis GWAS dataset. I benchmarked HapFM against widely used GWAS programs that marginally associate SNPs or haplotypes with phenotypes. The results clearly showed the advantages of HapFM over conventional GWAS algorithms in both mapping power and resolution of complex traits. In summary, HapFM serves as an alternative mapping algorithm to better understand the genetic architecture of complex traits and has the great potential of accelerating crop improvement.

Switching from a single reference genome to a pan-genome reference will be the next step for a more comprehensive and accurate discovery of plant diversity from the wild relatives. Although the situation is improving, many plant species still have a single reference genome due to the historical cost of generating a high-quality reference genome. As sequencing cost continues to decrease, an increasing number of studies have shown that a single reference genome is often inadequate in many genomic analyses resulting in biased or inaccurate results (Bayer et al., 2020; Della Coletta et al., 2021). A single reference genome is often insufficient in two dimensions, 1) as the divergence of the subject genome from the reference increases, there is an increase bias for mapping reads and calling variants containing reference alleles; and 2) extensive structural variation found in plants, especially in the form of presence/absence or copy number variation may result in entire segments of the subject genome being excluded from variant identification using a single reference.

A recent approach to overcome these limitations is to coalesce multiple divergent genomes into a composite pan-genome. Pan-genomes can incorporate major sequence and structural variation found in the species but absent in any single genome. This approach can alleviate the limitations associated with single reference (Bayer et al., 2020; Della Coletta et al., 2021; Golicz et al., 2016; Lei et al., 2021; Paten et al., 2017). For example, the tomato pan-genome uncovered the 4,873 new genes that were absent from the reference genome and identified a rare allele contributing to desirable tomato flavor (Gao et al., 2019b). The soybean pan-genome contains the complete sequence information from 26 representative accessions and identified 49.9% of the genes only present in a small set of individuals (Liu et al., 2020). Therefore, an optimized plant variant discovery workflow using a pan-genome reference might be the key to identifying untapped diversity in the species to help alleviate global food insecurity.

In Chapter 3, I provide a computational approach to the selection of pan-genome cohort aiming to maximize the haplotype representativeness in the pan-genome reference. The algorithm uses local haplotype information instead of global distance to select individuals for the cohort. HapPS consists of genome-wide block partition, representative haplotype cluster identification, and cohort selection by the Genetic Algorithm. HapPS prioritizes the haplotype representative of regions of high interest during the Genetic Algorithm step. Examples are high-diversity gene-coding regions. The benchmark study showed HapPS outperformed the global-distance-based method in five evaluation metrics, especially the average coverage of the high-diversity gene-overlapping blocks. The GO term analysis found the genes in these regions are enriched for environmental responding genes.

Finally, human selection for more resilient and improved crop varieties have been a primary focus in agriculture over the past thousands of years. With the unparallel technologies and resources, the integration of beneficial genetic diversity from wild species may help us combat the food security challenges and drive the next revolution in agriculture.

Chapter 1. Benchmarking Variant Identification Tools for Plant Diversity Discovery

Abstract

The ability to accurately and comprehensively identify genomic variations is critical for plant studies utilizing high-throughput sequencing. Most bioinformatics tools for processing next-generation sequencing data were originally developed and tested in human studies, raising questions as to their efficacy for plant research. A detailed evaluation of the entire variant calling pipeline, including alignment, variant calling, variant filtering, and imputation was performed on different programs using both simulated and real plant genomic datasets. A comparison of SOAP2, Bowtie2, and BWA-MEM found that BWA-MEM was consistently able to align the most reads with high accuracy, whereas Bowtie2 had the highest overall accuracy. Comparative results of GATK HaplotypCaller versus SAMtools mpileup indicated that the choice of variant caller affected precision and recall differentially depending on the levels of diversity, sequence coverage and genome complexity. A cross-reference experiment of *S. lycopersicum* and *S. pennellii* reference genomes revealed the inadequacy of single reference genome for variant discovery that includes distantly-related plant individuals. Machine-learning-based variant filtering strategy outperformed the traditional hard-cutoff strategy resulting in higher number of true positive variants and fewer false positive variants. A 2-step imputation method, which utilized a set of high-confidence SNPs as the reference panel, showed up to 60% higher accuracy than direct LD-based imputation. Programs in the variant discovery pipeline have different performance on plant genomic dataset. Choice of the programs is subjected to the goal of the study and available resources. This study serves as an important guiding information for plant biologists utilizing next-generation sequencing data for diversity characterization and crop improvement.

Introduction

Genomic technologies provide unprecedented opportunities to reveal the history of crop domestication, to discover novel genetic diversity, and to understand the genetic basis of economically important traits, collectively contributing to crop improvement and food security (Bevan et al., 2017). One of the most important steps in genomic analyses is the ability to accurately and comprehensively identify genetic variations. As sequencing cost continues to decrease, whole genome sequencing (WGS) strategies are increasingly employed for plant diversity and domestication studies. (Callaway, 2014; Hufford et al., 2012; Lin et al., 2014; Zhou et al., 2015). Accompanying improvements in sequencing technology is the need to not only improve but also better understand the algorithms that enable variant calling from sequencing data. Many of the algorithms used in the processing of sequencing data were originally developed and evaluated in human WGS studies yet are frequently used by plant genomic researchers (Chen et al., 2019b; Cheng et al., 2014; Li and Homer, 2010; Liu et al., 2013). The underlying assumption is that the performance of a given algorithm for human data will be similar for plant data, in spite of significant differences between the human and plant genomes.

The variant discovery pipeline for WGS dataset can be roughly divided into four steps: read mapping, variant calling, variant filtering, and imputation. Sequence aligners for the read mapping step can be grouped according to their indexing methodologies (Li and Homer, 2010). Programs such as Novoalign (http://www.novocraft.com) and GSNAP (Wu and Nacu, 2010) use hash tables indexing methods; whereas BWA (Li and Durbin, 2010), SOAP2 (Li et al., 2009b) and Bowtie2 (Langmead and Salzberg, 2012) use Burrows-Wheeler Transformation indexing algorithms. Variant calling programs can be categorized into alignment-based programs such as SAMtools (Li et al., 2009a) and FreeBayes (Garrison and Marth, 2012), and assembly-based programs, such as GATK HaplotypeCaller (Poplin et al., 2017) and FermiKit (Li, 2015). Variant filtering steps remove low-quality variants based on various quality metrics such as base quality, read depth, and mapping quality. The purpose of this step is to remove false positive variants while minimizing false negative variants, a source of "hidden diversity".

The basic filtering strategy, termed "hard-filtering" (De Summa et al., 2017), sets empirical cutoffs on quality metrics to eliminate false positive variants.

Over the past decade, extensive progress in human genomic studies has developed and applied machine-learning based variant filtering methods (Poplin et al., 2017) which uses adaptive cutoffs that adapt to a specific dataset, often by finding variants within the dataset that were previously identified with high confidence. The final step in variant discovery often employs imputation methods by leveraging external information to infer missing genotypes due to technical limitations. The standard way of imputation in human genomic studies utilizes a reference panel (Browning and Browning, 2016; Howie et al., 2009), where a previously identified set of haplotypes link missing variants with successfully genotyped variants. Many of these advanced methods have yet to be readily adopted by plant researchers. In some instances, there are clear obstacles to implementation, such as the lack of extensive plant haplotype panels of similar quality to the 1000 Genomes Project (Genomes Project et al., 2012) or HapMap (International HapMap et al., 2007). Though some species, such as maize (Bukowski et al., 2018) and rice (project, 2014), are rapidly building these resources. Even though both plants and human genomics share a similar computational workflow, the structure and composition of plant genomes pose unique challenges that are not present in humans. As a result, the evaluation of these emerging computational genomics technologies is urgently needed in agriculture.

A major challenge for crop genomics is the ability to accurately and comprehensively characterize genetic diversity in domesticated crops, diverse landraces, and wild crop relatives. Genetic diversity in plants can be much greater than that found human genomes. These sources of diversity, especially in the wild species, provide a reservoir of genetic variation for future crop improvement (Jacob et al., 2018; Migicovsky and Myles, 2017; Wulff and Moscou, 2014). For example, introgression from related wild species into domesticated tomato has been used to improve agronomic performance such as abiotic tolerance (Krause et al., 2018; Rambla et al., 2017; Zhang et al., 2018; Zhu et al., 2018). For example, a gene from a wild relative of bread wheat has been shown to confer resistance to one of the most destructive stem rust pathogen races, Ug99 (Periyannan et al., 2013). Characterizing these rich pools of diversity is an important challenge facing plant genomics because the regions containing this diversity may pose the most challenges for algorithms designed and optimized for human studies.

The second challenge for variant discovery in plant genomics is the quality of available reference genomes. The human reference genome has been in a constant state of improvement for decades (https://www.ncbi.nlm.nih.gov/grc/human). Once released, however, most plant reference genomes see little improvement, resulting in references that are less accurate and less complete than that found in humans. Other key challenges are the large amounts of repetitive sequences, structural variation and, in some crops, complex polyploid genomes (Michael and VanBuren, 2015; Schatz et al., 2012). Diversity may be underestimated because of presence-absence variations (PAV) that are common to most plant genomes (Wang et al., 2018b). The diverse nature of plant genomes together with low quality or incomplete reference assemblies can negatively affect read alignment and variant calling steps, leading to inaccurate genotypes and missing variants (Bevan et al., 2017; Huang and Han, 2014a; Morrell et al., 2011).

Here, we benchmarked the performance of programs that are commonly used for variant discovery in plant studies. The comparison included three highly-cited sequence aligners, BWA-MEM, Bowtie2 and SOAP2, and two popular variant callers, GATK HaplotypeCaller (GATK-HC) and SAMtools mpileup (SAMtools-mpileup) using domesticated tomatoes, wild relatives and simulated genomic datasets. We show that as diversity and genome complexity increased, the ability of these algorithms to identify variants varied. In addition, the inadequacy of a single reference genome was uncovered after a cross-reference comparison was performed. Finally, we evaluated the performance of machine learning based variant filtering method and reference panel assisted imputation methods on the high diversity plant datasets.

Results

Alignment program evaluation

The performance of three different aligners, BWA-MEM, Bowtie2, and SOAP2, was evaluated using Illumina paired-end read datasets from 52 domesticated tomato, 30 related wild relatives (Supplemental Table 1.1) (Tomato Genome Sequencing et al., 2014), and simulated genomic sequences from different crops. Mapping percentage, alignment accuracy, and processing time for each aligner were evaluated.

The ability to align reads to a domesticated tomato reference genome, *Solanum lycopersicum* (Tomato Genome, 2012), was assessed using default and tuned parameters on Bowtie2 (Bowtie2 and Bowtie2-tuned), SOAP2 (SOAP2 and SOAP2-tuned), and default parameters for BWA-MEM. Parameter tuning (see details in Methods) for Bowtie2 and SOAP2 was necessary to attempt to match the mapping percentage to the default used by BWA-MEM. BWA-MEM showed the highest alignment percentage, 99.54% and 95.95% in domesticated and wild relatives, respectively, while SOAP2 showed the lowest alignment percentage, 91.25% and 40.58%, respectively (Supplemental Table 1.2). In the domesticated tomato datasets, all of the five alignment settings resulted in more than 90% mapping percentage with standard deviation ranging from 0.34% to 3.77% (Figure 1.1A). Greater variation in mapping percentage existed when analyzing the sequences from wild species with standard deviation ranging from 1.91% to 24.25%. The mapping percentage in the wild tomato samples displayed a bimodal distribution (Figure 1.1A). The distribution of the group with higher alignment percentage contained wild species that were closely related to domesticated tomatoes, whereas the lower group contained distantly related wild species based on previous domestication and diversity studies (Lin et al., 2014; Strickler et al., 2015). Alignment percentage was found to be negatively correlated with the IBS distance of each sample to the *S. lycopersicum* reference genome (Figure 1.1B). When the sample was distantly related to the reference genome, BWA-MEM resulted in the highest mapping percentage and SOAP2 resulted in the lowest mapping percentage. In terms of processing time, SOAP2 was the fastest aligner in both domesticated and wild tomato datasets, and it was up to five times faster than the slowest alignment setting, Bowtie2-tuned (Supplemental Figure 1.1A).

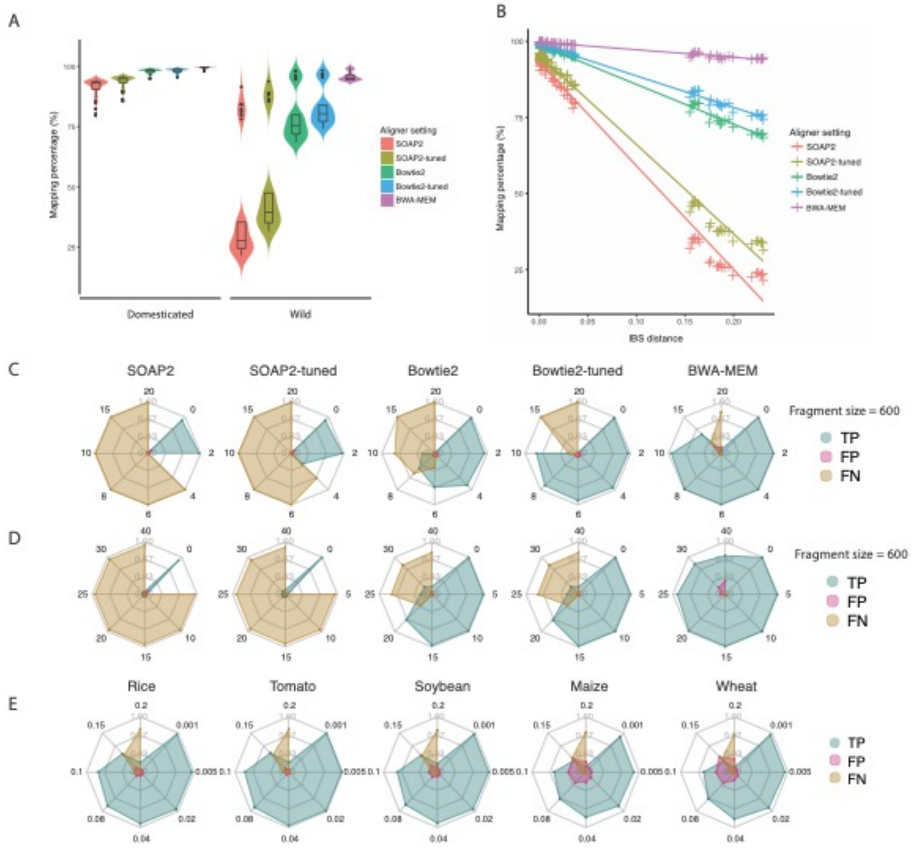


Figure 1.1. Aligner performance comparison using real and simulated plant genomic dataset

(A) Alignment percentage of five different aligner settings: SOAP2, SOAP2-tuned, Bowtie2, Bowtie2-tuned and BWA-MEM calculated for domesticated tomatoes and wild relatives. The width of violin plot is proportional to the density of the data. Boxplots inside violin plot indicate quantiles and outliers.

(B) Mapping percentage of each sample is shown relative to the IBS distance to the reference genome.

(C) Alignment accuracy of five aligner settings using simulated dataset with different number of SNPs per read and fixed 600 nt fragment size. Each axis represents the number of SNPs in the corresponding simulation. The blue color represents percentage of true positive (TP) alignments, pink color represents the percentage of false positive alignment (FP) and gold color represents the percentage of false negative (FN) alignments

(D) Alignment accuracy of five aligner settings using simulated dataset with different size of indels per read and fixed 600 nt fragment size. Each axis represents the size of indels in the corresponding simulation.

(E) Alignment accuracy of BWA-MEM on different crop species. Each axis represents different mutation rate which includes both SNP and INDEL mutations.

We next determined whether greater alignment percentage or shorter alignment time could result in tradeoffs on accuracy and sensitivity by using simulated datasets and calculating the ratio of true positive (TP), false positive (FP) and false negative (FN) alignments. Simulated datasets were derived from the reference genome by permuting fragment sizes, and number of SNPs or size of small indels per read. For all alignment methods, the ratio of FP alignment increased as more SNPs or indels were introduced per read (Figure 1.1C-D) when the fragment size was fixed at 600 nt. When the number of introduced SNPs was equal or less than 2, the average percent of FP alignments BWA-MEM, Bowtie2-tuned and SOAP2-tuned was 0.94%, 1.15% and 0.88%, respectively (Figure 1.1C). When the number of introduced SNPs was greater or equal to 4, the average FP alignment rate of BWA-MEM, Bowtie2-tuned increased to 6.41% and 2.54%, respectively, while SOAP2, and SOAP2-tuned were no longer able to find alignments. BWA-MEM was the only aligner that was capable of finding TP alignments with 15 SNPs per read with FP alignment rate of 18.26%. Similar results were also observed in the indel simulation experiment (Figure 1.1D). Only BWA-MEM was able to find TP alignments of reads with INDELs up to 40 nt in size at the cost of 26% false alignments. While differences in alignment percentages were observed, alignment length distributions were not found to differ for each aligner (Supplemental Figure 1.1B).

To indirectly determine the true vs false positive rates of BWA-MEM and Bowtie2 in real data, one million randomly selected reads from six samples (2 *S. lycopersicum*, 2 *S. pennellii* and 2 other wild relatives) were aligned to both *S. lycopersicum* and *Solanum pennellii* reference genomes (Bolger et al., 2014a). The positions of alignments with mapping quality (MQ) ≥40 were compared against the synteny map of the genome generated by nucmer (Kurtz et al., 2004). When the alignment position of read matched to the nucmer conversion of the *S. lycopersicum* coordinate to the *S. pennellii* coordinate, the read was considered to be syntenic. If the positions did not match, the read was considered non-syntenic. BWA-MEM was able to align approximately 4.22 times more reads per sample than Bowtie2 (Supplemental Table 1.3), but only 65.71% (*SD* ± 2.68%) of these alignments were considered as syntenic compared to 88.17% (*SD* ± 1.59%) of Bowtie2 alignments.

To extend the study to other crop species, simulated sequencing datasets were generated from rice, soybean, maize and wheat reference genomes by varying the mutation rate from 0.001 to 0.2 (Figure 1.1E). In these studies, both SNP and INDEL were included in the simulation. When the mutation rate is equal to or lower than 0.04, BWA-MEM was able to align at least 92% of the sequences correctly for rice, tomato and soybean, whereas it was only able to correctly align 81.5% and 82% of the sequences for maize and wheat, respectively. As mutation rate increased, difference in both true positive and false positive alignment was seen among different crops. On average, BWA-MEM was able to find 18.1%, 20.2% and 17.0% more true positive alignments in rice, tomato and soybean than in wheat and maize at mutation rate 0.08, 0.1, and 0.15, respectively. On the other hand, BWA-MEM was able to generate 18.8%, 22.5%, and 24.5% less false positive alignments in rice, tomato and soybean than in wheat and maize at mutation rate 0.08, 0.1, and 0.15, respectively.

Variant Calling Program Comparison

Four variant datasets were produced from the permutation of the aligners, Bowtie2-tuned, and BWA-MEM, and the variant callers SAMtools-mpileup and GATK-HC using 52 domesticated and 30 wild tomatoes. Results showed nearly a two-fold difference in the number of unfiltered SNPs ranging from 69.2M to 133.7M. A greater difference in the variant count in wild species was observed than that found in domesticated ones (Table 1.1). In domesticated species, dataset sizes ranged from 11.8M to 17.8M unfiltered SNPs, while in wild species it ranged from 66.4M to 128.3M. The primary determinant of variant count between datasets was whether Bowtie-2 or BWA-MEM was used. In domesticated species, 10.7M SNPs were commonly identified by different aligners and variant callers, and when BWA-MEM was used as the aligner, about 83% (14.7M) SNPs were identified by both GATK-HC and SAMtools-mpileup (Supplemental Figure 1.2A). In wild species, 59M SNPs were commonly identified by different aligners and variant callers, and when BWA-MEM was used as the aligners, about 84% (109.8M) SNPs were identified by both GATK-HC and SAMtools-mpileup (Supplemental Figure 2.1B). The inbreeding coefficient was calculated for each tomato individual, no significant difference (Wilcoxon rank sum test, *p*-value 0.47) was found between GATK-HC and SAMtools-mpileup identified SNP variants.

Table 1.1. Summary of SNPs identified by combinations of aligners and variant calling program

	Unfiltered SNPs				Filtered SNPs			
	Total	Domesticated tomatoes	Wild tomatoes	Common	Total	Domesticated tomatoes	Wild tomatoes	Common
BWA-MEM + GATK-HC	131,449,946	17,771,072	128,294,973	14,616,099	93,739,759	13,628,974	91,482,115	11,371,330
Bowtie2-tuned + GATK-HC	73,393,338	11,813,500	70,453,383	8,873,545	30,307,811	8,261,729	28,243,136	6,197,054
BWA-MEM + SAMtools-mpileup	133,734,683	17,268,821	130,886,221	14,420,359	80,709,232	10,366,835	78,727,565	8,385,168
Bowtie2-tuned + SAMtools-mpileup	69,219,499	12,390,916	66,416,422	9,587,839	46,436,709	8,832,598	44,626,459	7,022,348

To further evaluate the differences in the ability of identifying variants, both individual-level and population-level simulated datasets were generated with varied mutation rates, sequencing coverage and population size. In the simulated population-level datasets, evaluation was performed on both raw and filtered variants. In the comparison of raw variants, GATK-HC was able to identify more true SNPs at the cost of accuracy as sequencing coverage increased in diversity populations. At 5x and 10x coverages, SAMtools-mpileup was able to identify similar recall ratio with higher precision ratio than GATK in the low diversity population. When dealing with high diversity populations, GATK-HC always outperformed SAMtools-mpileup in both precision and recall aspects (Supplemental Figure 1.2C). When functional annotation was applied to each identified coding SNP, nearly identical percentages of missense, nonsense and silent SNPs were found between GATK-HC and SAMtools-mpileup (Supplemental Table 4). In the comparison of raw INDELs, GATK-HC always outperformed SAMtools-mpileup in terms of precision and recall in the low diversity population. In the high diversity populations, GATK-HC was able to identify greater number of true INDELs at the cost of accuracy (Supplemental Figure 1.2D). The true size of simulated INDELs ranged from -6 bp to 6 bp. The size of the raw INDELs identified by GATK ranged from -170 bp to 241bp, and size of the raw INDELs identified by SAMtools-mpileup ranged from -5 bp to 7 bp.

In the filtered SNP results, when the sequencing coverage is at 5x and 10x, GATK-HC provided a higher precision ratio in all coverage and diversity permutations without compensating the recall ratio (Figure 1.2A). In the 1x coverage simulation dataset, even though SAMtools-mpileup identified variants with lower precision ratio, it generated a higher recall ratio in the dataset. In the filtered INDEL results, GATK-HC always outperformed SAMtools-mpileup in terms of precision and recall ratio in the low diversity population. In the high diversity population, SAMtools-mpileup resulted in a higher precision ratio at the cost of a much lower recall ratio (Figure 1.2B). Noticeably, SAMtools-mpileup was only able to result in 3.08% and 1.61% recall ratio in the high diversity populations for SNPs and INDELs, respectively.

In the individual-level simulated dataset, a consistent pattern of trade-off between precision and recall was observed. SAMtools-mpileup was able to generate higher precision ratio for both SNPs and INDELs, however, GATK-HC was able to result in a higher recall ratio for both SNPs and INDELs as coverage and mutation rate increased in most case (Supplemental Figure 1.3A-D). Among four different crop species, rice, tomato and soybean has similar results in both variant calling programs. Nevertheless, results from simulated maize datasets showed lower precision and recall ratios. Noticeably, when the mutation rate is at 0.1 and 0.15, both variant calling programs resulted lower precision ratio for SNP detection as coverage increased. Maize datasets had the largest magnitude of reduction in precision whereas other crop species resulted similar reduction.

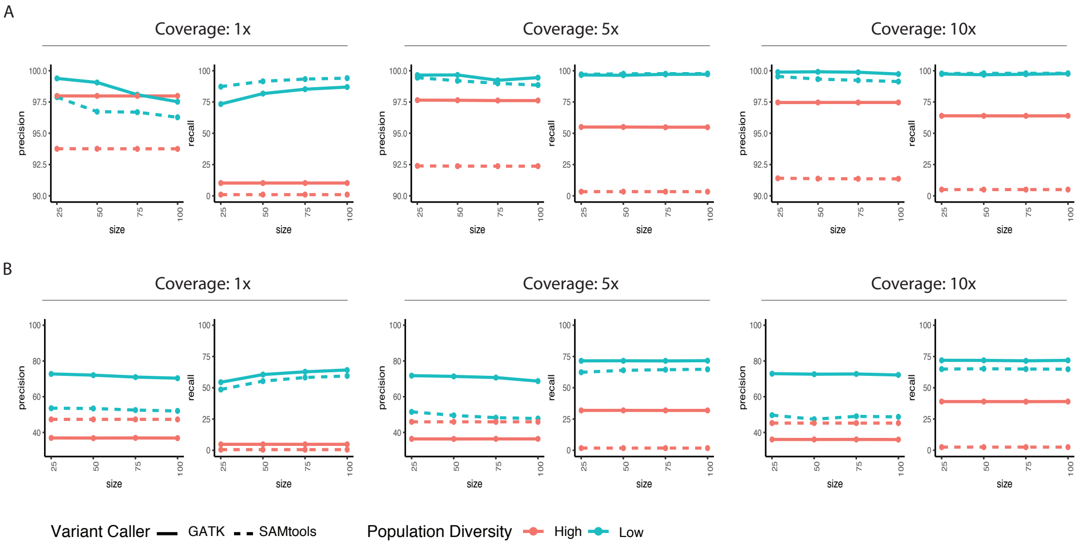


Figure 1.2. Evaluation of variant calling programs using simulated plant genomic datasets

(A) The comparison of the performance of GATK-HC and SAMtools-mpileup on filtered SNPs at different coverages, population diversity and population size.

(B) The comparison of the performance of GATK-HC and SAMtools-mpileup on filtered INDELs at different coverages, population diversity and population size.

Wild reference genome alignment and variant calling

The large increase in the number of SNPs in wild samples was expected due to both greater distance from the domesticated reference genome and increased diversity relative to the domesticated samples. Expectedly, as the distance from the reference genome increased, a greater proportion of reads was unmapped. The variants in these unmapped reads, especially in the wild species, could represent "missing diversity". To test this hypothesis, we evaluated how variants discovery in these 82 tomato samples were changed by mapping reads to a wild reference genome (*S. pennellii*) (Bolger et al., 2014a).

The read alignment to the *S. pennellii* reference was performed under identical settings as above. As previously seen, BWA-MEM showed the highest mapping percentage and SOAP2 showed the lowest (Figure 1.3A). In general, mapping percentage in domesticated and wild tomato groups were similar regardless of aligner settings used (Figure 1.3A). The single outlier with high alignment percentage was a *S. pennellii* sample with an alignment of 95.13% (or 99.69%) as opposed to 34.22% (or 94.87%) against the *S. lycopersicum* reference using SOAP2-tuned (or BWA-MEM). Interestingly, the 82 samples, except for the *S. pennellii* sample, had similar IBS distances to the reference genome. As with the *S. lycopersicum* reference, alignment percentage to the *S. pennellii* reference was inversely proportional to IBS distance to the reference genome (Figure 1.3B), suggesting this relationship was independent of reference genome used.

To investigate how diversity estimation varied by reference genome, reads from randomly selected eight domesticated tomatoes and eight wild relative accessions were aligned to the *S. pennellii* reference. Alignment to the *S. pennellii* reference genome generated a total of 96,712,749 unfiltered SNPs and 59,944,499 filtered SNPs, while a total of 77,718,102 raw SNPs and 53,036,666 filtered SNPs were identified using the *S. lycopersicum* reference genome. Compared to using the *S. lycopersicum* reference genome, significantly more SNPs (Two-sample T-test, *p*-value = $2.3*10^{-10}$) were identified from 8 domesticated tomato samples when *S. pennellii* reference genome was used for variant discovery (Supplemental Figure 1.4A).

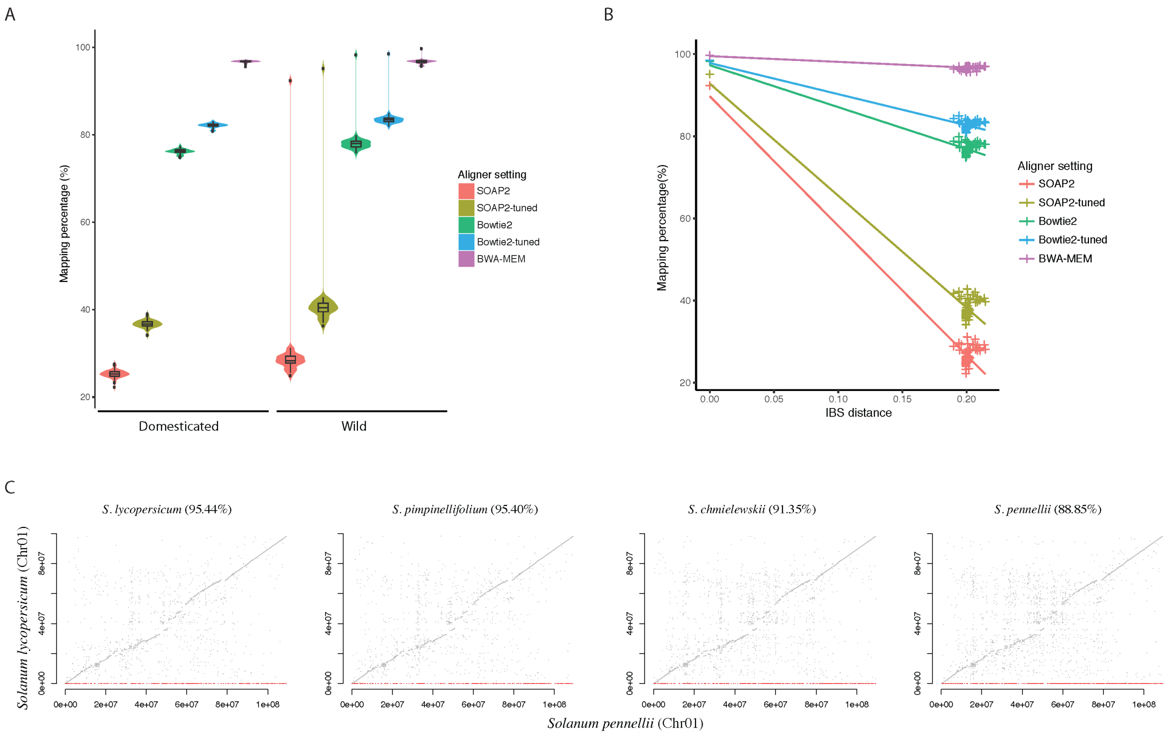


Figure 1.3. Alignment and variant calling using a wild reference *S. pennellii* genome

(A) Alignment percentage of five different aligner settings: SOAP2, SOAP2-tuned, Bowtie2, Bowtie2-tuned and BWA-MEM calculated for domesticated tomatoes and wild relatives using the *S. pennellii* reference genome. The width of violin plot is proportional to the density of the data, and boxplot is plotted inside violin plot showing quantiles and outliers.

(B) Mapping percentage of samples for different aligner setting. The mapping percentages are relative to the IBS distance to the reference genome

(C) SNP identification of four tomato samples was performed in chromosome 1 in *S. pennellii* reference genome. The corresponding physical positions of SNPs in the *S. lycopersicum* reference was plotted. The grey dots represented the SNPs that were able to be located at the corresponding positions in *S. lycopersicum* genome, red dots represented the SNPs that were unable to be located to corresponding positions in *S. lycopersicum* genome. The percentage of corresponding SNPs are written next to the species name.

To further investigate the source of this additional variation, a cross-reference comparison was performed between SNPs identified using *S. pennellii* and *S. lycopersicum* reference genomes. One hundred nucleotides of DNA sequence flanking each filtered SNP identified using one reference genome was aligned to the other reference. Results in the Figure 1.3C showed that majority of the filtered SNPs identified in the *S. pennellii* located on the synteny path of *S. lycopersicum* genome. In the *S. lycopersicum* sample, and similarly, majority filtered SNPs identified using *S. pennellii* reference were located on the synteny path of *S. lycopersicum* genome. This result indicated that using *S. pennellii* reference genome, we were able to identify SNPs that were fixed in the *S. lycopersicum* domesticated varieties.

Since these SNPs were fixed in *S. lycopersicum*, they would not have been identified from alignment to the *S. lycopersicum* reference. Outside of these fixed SNPs in the domesticated species, 4.55% of flanking sequences of SNPs identified using *S. pennellii* genome in chromosome 1 could not be mapped to the *S. lycopersicum* reference. Similarly, 11.15 % of the flanking sequences of SNPs identified in the *S. pennellii* sample using the *S. pennellii* genome were not found in the *S. lycopersicum* genome (Figure 1.3C). Switching to the domesticated reference genome, 7.13% of the downstream sequences of SNPs identified in a *S. lycopersicum* sample using *S. lycopersicum* genome could not be found in the *S. pennellii* genome (Supplemental Figure 1.4B). These results indicated that a great portion of variation in the wild species would be missed if a single domesticated genome was used as the reference, and vice versa.

Hard-filtering and machine-learning based variant filtering

Variant filtering is required to minimize both false positive and negative genotype calls. Comparisons were made between three variant filtering methods: setting empirical hard-cutoffs (HARD) on metrics such as read depth, strand bias, and variant quality and so on, a newly implemented machine-learning based (ML) variant filtering (Poplin et al., 2017), and a combination between HARD and ML (COMBINED) filtering. Filtered datasets generated from the 602 WGS tomato datasets, including a wide range of domesticated and wild tomato samples (Zhu et al., 2018), were analyzed. A training dataset of 8401 markers from SolCap was used for the training phase of ML (Sim et al., 2012). The SolCap is a high confidence dataset consisting of verified markers previously used in genetic studies. In the COMBINED method, the HARD filters were first applied to SolCap to remove low-confidence markers and yield a training set of 7,633 variants. Results indicated that the HARD-filtered method retained the fewest SNPs (94.2M), which was 26.3% and 7.1% fewer than ML-filtered (127.8M) and COMBINED-filtered (101.4M) datasets, respectively (Supplemental Table 1.5). SNPs in the first 10 million bases in Chromosome 1 (Supplemental Table 1.5) were cross-compared between the three datasets. 70% of SNPs in this segment were shared among all three filtered datasets (Supplemental Figure 1.5A), while each dataset had a subset of unique variants.

Two methods were used to indirectly infer the quality of filtered datasets: recapitulation of diversity estimates generated by a "gold standard" set of 22,336,965 SNPs (See details in Methods) in the form of PCA (Supplemental Figure 1.5B) and IBS analyses (Supplemental Figure 1.5C), and calculation of LD decay distance for each filtered dataset. SNPs identified by all three filtering methods were removed for this analysis so that the efficacy of each method could be evaluated independently. The underlying assumption of these analyses is that true diversity would recapitulate the known population structure, whereas the population structure would begin to break down as the number of artifacts increased. Using the "gold standard" variant dataset, samples were grouped into four clusters based on PCA and IBS results. All three filtering methods were able to resolve Cluster 1 and Cluster 4, whereas the HARD and ML filtering methods were not able to clearly resolve Cluster 2 from Cluster 3 (Figure 1.4A-B). In contrast, the COMBINED filtering method was able to identify all four original clusters to reconstruct the population structure of 82 Solanum genomes (Figure 1.4C).

Next, the contribution of false positive SNPs in each filtered dataset was evaluated by calculating the rate of LD decay. The assumption was that false positive SNPs were random noise that would be found not in LD with nearby SNPs. Therefore, the apparent rate of LD decay in a dataset would increase as the number of false positives increased. As predicted, a greater rate of LD decay was found in all three filtered datasets than that found in the high-confidence dataset. Of the three filtered datasets, the COMBINED method, however, had the lowest rate of LD decay (Figure 1.4D) approximating the rate of LD decay seen in the high-confidence SNP dataset.

To quantitively measure the difference between hard filtering and machine-learning based filtering, simulated datasets with varied population size, mutation rate and sequencing coverage were generated (Supplemental Figure 1.6A-B). In the simulation analysis, 30% of the simulated gold standard variants were used as the training dataset, and no hard-filtering was performed on the training dataset. In the low diversity population datasets, machine-learning based SNP filtering always outperformed hard SNP filtering by 7.38% and 14.14% on average for precision and recall ratio, respectively. In terms of INDEL filtering in the low diversity dataset, machine learning based filtering and hard filtering resulted comparable precision results, however, machine learning based filtering was able to result 12.49% higher recall ratio than hard filtering. In the high diversity population, SNP and INDEL had similar results from different filtering methods. Minor difference was observed in the recall ratio between machine-learning based and hard filtering. No difference was found in the precision ratio between machine-learning based and hard filtering in the high diversity population.

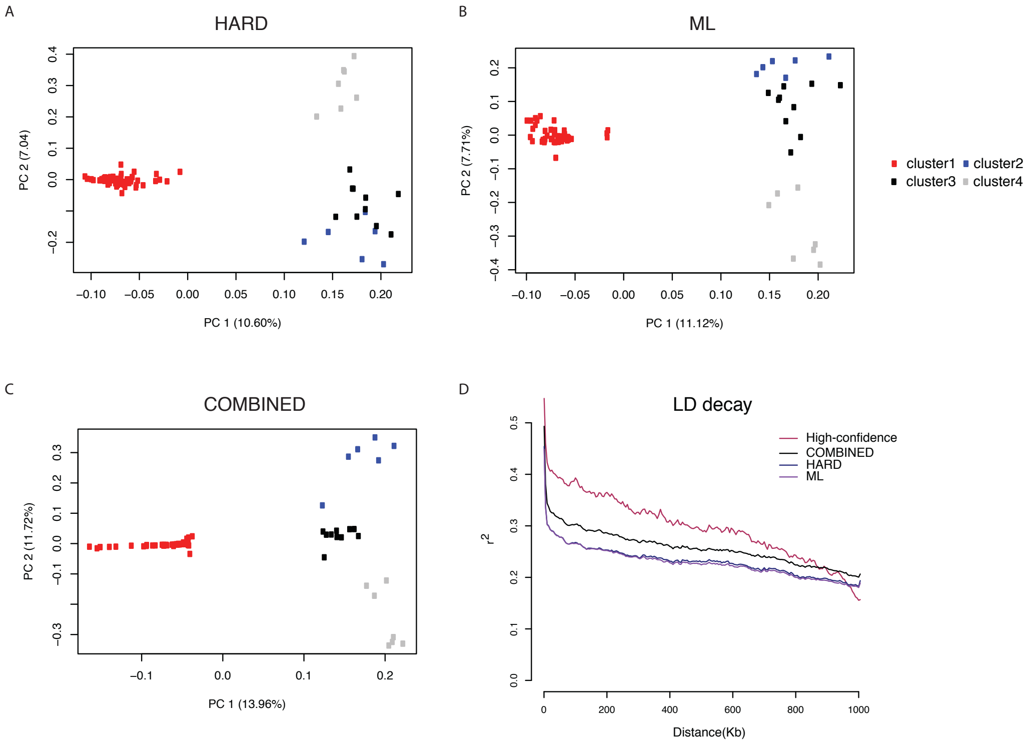


Figure 1.4. Comparison between three variant filtering methods using PCA and LD decay to estimate false positive and false negative ratios.

(A) Unshared hard-filtered SNPs were not able to clearly separate cluster 2 and 3

(B) Unshared machine learning SNPs were not able to clearly separate cluster 2 and 3

(C) Unshared combined SNPs were able to clearly separate 4 clusters.

(D) Comparison of LD decay among four sets of SNPs

Two-step imputation method

Missing genotypes, possibly due technical limitations, are commonly resolved via imputation. In human studies, standard imputation methods leverage linkage disequilibrium (LD) and reference panels (Das et al., 2016). Beagle 4.1 is a commonly used imputation algorithm in plant studies that can function with or without a reference panel. To determine the importance of a reference panel for SNP imputation, both LD-based and reference panel-assisted imputation were applied to several datasets. A reference panel of 22,336,965 high-confidence, phased SNPs was generated from 82 high coverage (30x) WGS tomato datasets. Imputation results were compared between the two methods. In the first method, missing SNPs were imputed without a reference panel. In the second method, imputation was performed in two steps: in the first step a reference panel was used to impute missing calls only for missing reference variants; and then a second step was employed to impute the remaining missing, non-reference SNPs. Samples were placed in four groups and varying percentages of high confidence genotypes were masked to act as “missing” data (See details in Methods). The concordance (*r*-squared) between the original masked and imputed genotypes was calculated to estimate imputation accuracy.

Results showed that no difference between LD-based and 2-step imputation was observed in 100 domesticated (DOM) tomato samples (Figure 1.5A) or the 50 *Solanum pimpinellifolium* (PIM) samples (Supplemental Figure 1.7A) datasets. In the dataset of 200 randomly selected tomato samples (RANDOM), at 47% missing data, a 4% difference was observed (Supplemental Figure 1.7B). When the parameter of missing percentage was set at 72%, 2-step imputation methods showed 60% higher accuracy than LD-based imputation in the dataset of 36 wild tomato species (WILD) (Figure 1.5B). High LD between SNPs may reduce the need for a reference panel in imputation. The calculated LD decay for each dataset showed that DOM had the slowest LD decay and WILD had the fastest LD decay (Supplemental Figure 1.7C). Due to the fact that limited samples of wild tomato were available, the number of samples we used in the simulation in DOM (100) was also considerably higher than that in WILD (36). As such, considerably more information was present in the DOM dataset for imputation, as opposed to the WILD dataset which not only had a smaller number of samples but also contained multiple species. To determine if LD continued to be sufficient for imputation in small domesticated panels when the amount of missing data was considerable, 15 randomly selected domesticated tomato samples that were also included in the reference panel (15-DOM-REF) had up to 85% of their genotypes masked. Both methods were applied to the 15-DOM-REF dataset. The results showed two-step imputation was 9.25 times more accurate than Beagle v4.1 direct imputation by when the missing percentage was 85% (Figure 1.5C).

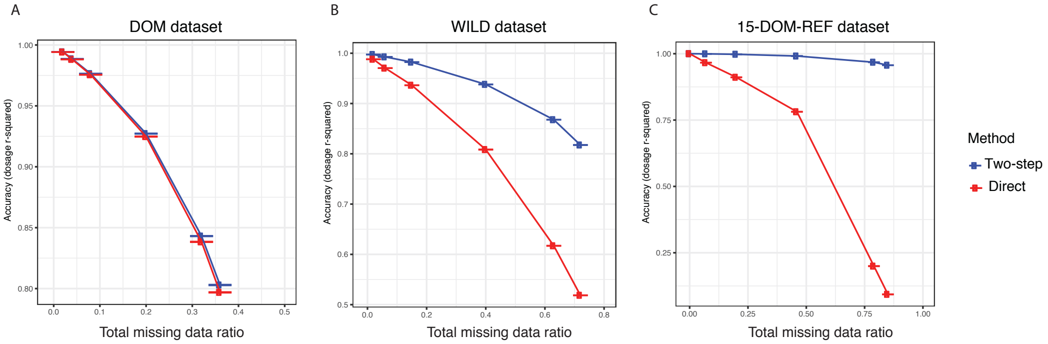


Figure 1.5. Comparison of imputation accuracy using direct imputation and 2-step imputation methods

(A) Imputation accuracy using direct imputation and 2-step imputation relative to missing SNPs in 100 domesticated tomato samples

(B) Imputation accuracy using direct imputation and 2-step imputation relative to missing SNPs using 36 wild samples

(C) Imputation accuracy using direct imputation and 2-step imputation relative to missing SNPs using 15 domesticated samples used in reference panel

Discussion

The ability to accurately and comprehensively identify genetic variation is a critical step for studying diversity, trait mapping and breeding in plant genomics. Many plant studies involve high levels of genetic diversity and, in some instances, incorporating distantly related varieties and wild relatives. Neither of these conditions are common in human studies, and as such pipelines designed and evaluated on humans may perform differently than expected. Therefore, we evaluated programs that are commonly used by plant genomic studies on SNP discovery steps including read alignment, variant calling, variant filtering and missing data imputation in the context of plant diversity discovery

One of the first computational steps in the variant discovery pipeline is the alignment of reads to a suitable reference genome. Previous aligner evaluation studies have been performed using either human or microbial genomic datasets (Shang et al., 2014; Thankaswamy-Kosalai et al., 2017), which may not represent the levels or types of diversity expected in plant studies. We performed alignment using both real and simulated tomato datasets and found that different aligners were very different in their tolerance of sequence variation in paired-end reads. BWA-MEM outperformed four other alignment settings in mapping percentage while still being able to maintain high mapping accuracy. Neither SOAP2 nor Bowtie2 was able to align as many reads, even after optimizing their settings to account for increased variation. In this study, we chose not to tuned BWA-MEM mostly because the mapping percentage was high with the default settings and there is no obvious parameter such as numbers of mismatches allowed, or fragment size as found in Bowtie2 or SOAP2. Besides, many program users, especially non-experts in bioinformatics, may stay with the default settings of programs.

BWA-MEM's increased sensitivity may come at a cost in that, as the number of SNPs or size of INDELs per read increased, the false positive rate also became slightly higher than that of Bowtie2-tuned (Figure 1.1C-D). The increased number of false positive alignments may, in turn, result in erroneous variant identifications. Nevertheless, given the relatively high sensitivity and accuracy of BWA-MEM, our results indicate that under most circumstances it is probably the most suitable algorithm for read mapping for plant datasets, especially when distantly related samples are included in the analysis. If high accuracy at the cost of less sensitivity is desired, Bowtie2 may be the better choice. Although SOAP2 was the fastest aligner tested, its difficulty in aligning reads with high variance from the reference genome make it unsuitable for studies where significant levels of genomic diversity may be present.

The next step in an analysis pipeline is variant calling. Comparisons between aligner-variant caller combinations indicated that the alignment algorithm had a greater impact on the number of variants discovered than the variant caller used. For a given aligner, SAMtools-mpileup and GATK-HC had similar results in the total number of SNP identified in the real tomato genomic dataset. This further emphasizes the importance of selecting an aligner appropriate to the goals of the experiment, especially when high diversity samples such as wild relatives and related species are included in the study. According to the simulation results, GATK-HC was able to identify more true positive variants at higher precision ratio in most population-based variant discovery cases. Especially in the high diversity population simulation, GATK-HC was more preferred than SAMtools-mpileup because SAMtools-mpileup resulted very low recall ratio in both SNP and INDEL detection. In the simulation analysis, the size of INDELs identified by GATK-HC had larger range than those identified by SAMtools-mpileup and ground truth INDELs, which partially explained why GATK-HC had lower precision in the INDEL identification than SAMtools-mpileup. One of the possible explanations is that GATK-HC performs local-assembly to identify the haplotypes whereas SAMtools-mpileup only utilizes read alignments. Plant genomes, in general, are rich in repetitive sequences which are difficult to assemble correctly using short reads. Therefore, the local assembly strategy taken by GATK-HC will not only identify true variants, but also generate false positive variants, INDELs especially. Consistent with a previous research (Clevenger et al., 2015), SAMtools-mpileup resulted higher precision value than GATK-HC for variant identification in the individual-based genotyping. However, the high precision of SAMtools-mpileup is at the trade-off of low recall value.

In general, we recommend GATK-HC for variant calling and filtering for several reasons. First of all, GATK-HC outperformed SAMtools-mpileup in most of our situation tests resulting a higher precision and recall ratio for SNP and INDEL detection. Second, GATK-HC allows rapid incorporation of multiple samples into a dataset without needing to recall genotypes for all samples, even previously genotyped ones, from aligned reads by using the GVCF system. This saves considerable time and computational expense when adding samples to a dataset. The third reason to recommend GATK-HC is that it supports multi-thread processing which is not available in the SAMtools-mpileup. Taking the advantage of high-performance clusters, multi-thread feature can significantly save processing time especially for large studies. Finally, the GATK package supports sophisticated machine learning based variant filtering (VQSR) which showed superior performance than empirical hard cutoffs. We did, however, find situations that SAMtools-mpileup is more preferable depending on the goal of the study. For example, for a low diversity population with very low sequencing coverage (1x), SAMtools-mpileup was able to identify more true SNPs than GATK-HC but at the cost of lower precision. If the purpose of the experiment is to identify as many true positive SNP as possible, then SAMtools-mpileup could be used in this particular situation. Another situation that SAMtools-mpileup may be preferable is identifying SNP from a closely related sample. According to the simulation results from single samples, SAMtools-mpileup resulted slightly higher precision and recall values than GATK-HC results when the mutation rate was lower than 0.05. If the experiment aims at charactering SNPs in a line that is closely related to the reference genome, SAMtools-mpileup could be used in this particular situation.

Variant filtering is the third step in a diversity assessment pipeline. Three approaches to this were evaluated: hard filters of various quality metrics, machine learning as implemented in GATK (VQSR), and a combined approach. The combined approach which utilized hard filtered SolCap markers as the training dataset showed significant improvements over other variant filtering methods. According to the PCA plots (Figure 1.3A-C) and LD decay figure (Figure 1.3D), the combined method was able to generate more true positives, with fewer false negative SNPs and fewer false positive SNPs when an appropriate training dataset was used. This indicates that machine-based learning methods may be better suited at identifying true positives and eliminating false positive SNPs than empirical hard-filtering. The difference in the results of combined and VQSR suggested the importance of the training dataset. The machine learning model will learn from errors in the training dataset that might contribute to false positive variants. The downside of machine-learning-based filtering is that its implementation is complicated and requires experimentally validated (high-confidence) training set. In human studies, this information can be obtained from numerous genomic resources such as HapMap, the 1000 Genomes Project and omni SNP array datasets. Only in few major crops, such as maize (Bukowski et al., 2018), rice (Thomson et al., 2017) and soybean (Song et al., 2013) are these resources available. Similar conclusions were found from the simulation tests. According to the simulation results, VQSR outperformed hard filtering in general. Nevertheless, only minor difference was found when the simulated population had high diversity for both SNP and INDEL filtering suggesting the quality metrics used by VQSR may not be sophisticated enough to differentiate true variants from false positive variants. This also indicates new quality metrics may be necessary, especially for the genomic regions that can be hyper-variable.

The final step in the variant discovery pipeline is imputation. Reference panels are routinely employed in human studies, but these have not been routinely employed in plant genomics. To evaluate the importance of a reference panel for imputation, Beagle v4.1 (Browning and Browning, 2016) was used to impute masked genotypes in four sample group without the use of a reference panel and with a reference panel in a two-step process where SNPs contained in a reference panel were first imputed, and then imputation was extended to the entire dataset. Our results showed that the two-step imputation method was able to utilize a *de novo* reference panel of SNPs generated from high coverage sequencing data to assist imputation in the low coverage samples. Results from these studies indicated that the two-step imputation method was superior to the LD-based imputation method in sample groups that contained wild species. In addition, even for closely related samples, a certain number of samples must be present for LD-based imputation to produce valid results. Further, if there are insufficient samples, a reference panel may be required (Figure 1.4C). The tradeoff was that 2-step imputation doubled the running time and would incorrectly impute missing SNPs which were not due to technical issues but because of structural variations. Therefore, care must be taken not to introduce false positive since presence-absence variations are common in plants. These genomic regions could be identified prior to imputation to avoid this pitfall.

The effect of presence-absence variation on identifying missing genetic diversity is a special concern in studies that include high diversity samples. This issue can be seen from the results of the cross-reference experiment. Up to 11.15% of the variations identified using the wild reference could not be mapped back to the domesticated *S. lycopersicum* genome, and vice versa. These results indicated the inadequacy of single reference genome for comprehensive variant discovery. It also indicated that employing multiple reference genomes could identify additional sources of diversity that went undetected when using a single reference. These results have implications for the utility of pan-genomes. Multiple references or pan genomes would likely increase the detection of "missing diversity" that is due primarily to PAV between samples. Moreover, using a distantly related reference genome may allow the detection of SNPs that would be undetected using a closely related reference genome. These species-specific, fixed variants have implications in the evolutionary history of plant species such as domestication events. To date, several crop pan-genomes have been reported (Gao et al., 2019a; Hubner et al., 2019; Yu et al., 2019) that show significant amount of structural variations in the genome. Pan-genomes resources should be included into the diversity discovery pipeline in the future. Yet, one of the potential issues that will need to be addressed is that pan-genome assembled from diverse individuals may introduce more assembly errors than a single reference assembly. The quality of the reference genome will impact variant discovery because bioinformatic tools assume the reference genome is correct and only identify differences accordingly. Moreover, the level of heterozygosity of the reference introduced by the pangenome may require additional fine-tuned parameters (Kim et al., 2014). The most effective approach of utilizing a pan-genome reference will be a subject of future investigation.

Conclusion

In conclusion, we found that BWA-MEM was better overall at detecting more true-positive alignments, especially in distantly related samples, while Bowtie2 was better at minimizing the incorrect alignments. Incorporating multiple reference genomes gave a more complete picture of variations, especially when the samples showed considerable presence-absence variation. For filtering, the optimal approach found in our test was to incorporate a combination of machine learning and hard filtering, in which a set of "known" SNPs was used as the training set for machine learning. This requires a panel of known, high-quality SNPs however, which may be unavailable for many plant species. Finally, the importance of high-quality reference panels was emphasized during the imputation step especially when genotype imputation was challenging due to small LD blocks or not enough samples. Above all, the computational pipeline to discover variation from plant sequencing data will depend upon the diversity of the datasets, whether the goals of the experiment benefit from higher sensitivity or accuracy, the depth of sequence coverage, and the availability of external resources such as reference panels and gold-standard SNPs.

Materials & Methods

Simulated multi-species genomic dataset and real tomato genomic dataset

We used publicly available 602 WGS datasets representing 514 domesticated and 88 related wild species of tomato. The data were retrieved from the NCBI BioProjects under accession PRJNA259308, PRJNA353161 and PRJEB5235. The raw sequence data was quality trimmed using Trimmomatic (version 0.36) (Bolger et al., 2014b) with the options ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:8:TRUE SLIDINGWINDOW:4:20 LEADING:5 TRAILING:5 MINLEN:36. PCR duplicates were removed using Picard MarkDuplicates (version 2.14.1)) (http://broadinstitute.github.io/picard/). Simulated tomato sequencing reads were generated from the *S. lycopersicum* using a custom Python script was used to introduce from 0-20 SNPs per read, fragment sizes ranging from 200-10000 nt, and INDELs ranging from 0-40 nt. In order to evaluate the performance of BWA-MEM on multiple crop species, simulation of the Illumina sequencing reads was also performed on rice, soybean, tomato, maize and wheat using mason (version 2.0.9) (Reinert et al., 2017). The mutation rate including SNPs and INDELs was simulated at 0.001, 0.005, 0.02, 0.04, 0.08, 0.1, and 0.15. The proportion of the SNPs and INDELs were 0.85 and 0.15, respectively. Sequencing error was modeled as the default settings.

Evaluation of read alignment programs

Different aligners were evaluated using high-coverage datasets from PRJEB5235 and simulated datasets. BWA-MEM (version 0.7.17-r1188), SOAP2 (version 2.21), SOAP2-tuned, Bowtie2 (version 2.3.3.1) and Bowtie2-tuned were tested. SOAP2-tuned was used with the following options: -m 100 -x 888 -s 35 -l 32 -v 3 (Zhu et al., 2018). Bowtie2-tuned was used with the following options: --very-sensitive -N 1 -I 100 -X 888. To determine mapping percentages, these five aligner settings were used to align one million reads that were randomly selected from high coverage genomes from 52 domesticated and 30 wild relative samples. The IBS (Identity-By-State) distance was calculated using SNPrelate (version 1.16.0) (Zheng et al., 2012). The true positive alignments ratio was calculated by comparing the known ground truth location and aligned location. BWA-MEM was also evaluated on multiple crop species with a mixture of SNPs and INDELs in the simulated datasets.

SNP discovery comparison

Eighty-two high-coverage datasets from PRJEB5235 was used for SNP discovery comparisons. SNPs were called with SAMtools-mpileup (version 1.9) and GATK-HC (version 3.8-0-ge9d806836) using BWA-MEM and Bowtie2-tuned alignments. In GATK, variants were firstly identified by HaplotypeCaller using the option --emitRefConfidence GVCF, and then joint genotyping was performed using GenotypeGVCFs. In SAMtools-mpileup, genotyping was done in one step and the option -C 50 was used as recommended in the manual. Only polymorphic SNPs were used as data for the Venn diagram. Simulated datasets with known variants were generated for tomato, rice, soybean, maize using mason. Each crop species was simulated at different coverages (5x, 15x, 30x, and 50x) and mutation rates (0.001, 0.01, 0.05, 0.1, 0.15). In addition to individual simulated datasets, population-level simulated datasets were also generated with varied diversity (low diversity: 0.001 mutation rate and high diversity: 0.1), population size (25, 50, 75, and 100) and sequencing coverage (1x, 5x, and 10x). SAMtools-mpileup and GATK-HC were evaluated on both individual and population simulated datasets by comparing the precision and recall ratios. The functional annotations of the variants were predicted by snpEff (version 4.3) (Cingolani et al., 2012).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Imputation algorithm comparison

Beagle v4.1 (Browning and Browning, 2016) direct imputation and 2-step imputation method were compared using 602 tomato genomes. The raw SNPs were called using BWA-MEM and GATK-HC pipeline, and then hard filtered using GATK recommended options: "QD <2.0 || FS > 60.0 || MQ < 40.0 || SOR > 3.0". The high-confidence set of SNPs for the 2-step imputation was identified from 82 high-coverage dataset using BWA-MEM and GATK-HC. GATK hard-filtering and VCFtools (Danecek et al., 2011) with options: --missing 1 and --mac 2. SNPs with heterozygosity above 20% were removed. Beagle v4.1 was used to phase the high-confidence set of SNPs. The comparison was performed on four groups of samples: Two hundred random tomato and wild samples (RANDOM), 100 domesticated tomato samples (DOM), 50 *Solanum pimpinellifolium* samples (PIM), and 36 distantly related wild species (WILD). The one hundred domesticated samples from PRJNA353161 only, 15 DOM-REF samples from PRJEB5235 only, 50 PIM samples and 36 WILD samples were randomly selected for generating simulated datasets. Polymorphic SNPs in each dataset were randomly masked using a custom Python script if there were more than 7 reads supporting the genotypes. Both Beagle v4.1 and 2-step imputation methods were used to impute missing genotypes in five simulated datasets. The concordance R^2 ratio between genotyped and imputed values were calculated as imputation accuracy using BCFtools (Danecek and McCarthy, 2017).

Variant filtering algorithms comparison

The 602 tomato datasets were used to generate raw SNPs using BWA-MEM and GATK-HC pipeline. Hard-filtered, machine-learning based and combined filtering methods were individually applied to the raw dataset. The parameters used for hard-filtering included QualByDepth (QD < 2), FisherStrand (FS > 60), RMSMappingQuality (MQ < 40) and StrandOddsRatio (SOR > 3.0), which was suggested by the GATK hard filtering tutorial (https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set). For INDELs, hard filtering was performed using "QD <2 || FS > 200 || ReadPosRankSum < -20", as suggested by the GATK tutorial. The machine learning based methods, for both SNPs and INDELs, followed the GATK Best Practice Workflow (https://software.broadinstitute.org/gatk/documentation/article.php?id=2805). To summarize, the first step was to build a variant recalibration model using the program VariantRecalibrator. In the real tomato genomic dataset, SolCap and filtered SolCap markers were used as the training dataset with prior likelihood set to 90% and 95%, respectively. In the simulated dataset, 30% of the simulated gold standard variants were used as the training dataset with the prior likelihood set to 95%. All the annotations generated by GATK-HC including coverage, coverage by depth, FisherStrand, StrandOddsRatio, MappingQualityRankSumTest, ReadPosRankSumTest, RMSMappingQuality and InbreedingCoeff, were used to build the recalibration model. The second step is to apply the recalibration model to variants using the program ApplyRecalibration with the option --ts_filter_level 99.9. Polymorphic SNPs in the first 10 million base pairs of Chromosome 1 were selected to test the performance of different filtering methods. PCA was performed using SNPrelate after LD pruning ($R^2 > 0.2$). LD decay was calculated using the PopLDdecay package (Zhang et al., 2019) with default parameters.

Chapter 2. Causal Haplotype Block Identification in Plant Genome-Wide Association Studies

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.

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 (Visscher et al., 2017). For example, the identification of an androgen receptor (AR) gene through GWAS led to the development of therapeutic drugs for patients with prostate cancer (Farashi et al., 2019). GWAS methods have also been used in plant studies to identify the genetic basis of certain agronomic traits (reviewed by (Cortes et al., 2021)). There have been many successful applications including the identification of *OsSPY* for plant architecture in rice (Yano et al., 2019), metabolic genes for tomato flavor (Tieman et al., 2017), and *ZmFBL41* for blight resistance in maize (Li et al., 2019). 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 (Cortes et al., 2021; Huang and Han, 2014b) (Xiao et al., 2017; Zhou and Huang, 2019). 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 (Cortes et al., 2021; Zhou and Huang, 2019). 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 (Segura et al., 2012), SUPER (Wang et al., 2014b) and FarmCPU (Liu et al., 2016) 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 (Cortes et al., 2021). In human GWAS studies, SNP-set based GWAS method, SMMAT (Chen et al., 2019a) 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 (Hamazaki and Iwata, 2020) and FH-GWAS (Liu et al., 2019), 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 (Badouin et al., 2017; Lin et al., 2014; Zhou et al., 2015). 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 (Cortes et al., 2021) (Ingvarsson and Street, 2011), 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 (Li et al., 2020; Wang et al., 2018a; Wang et al., 2016). 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 (Schaid et al., 2018). 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 (Goktay et al., 2021), rice (Fuentes et al., 2019), tomato (Alonge et al., 2020), soybean (Anderson et al., 2014), maize (Yang et al., 2019) to name a few. Similarly, the availability of transcriptomic datasets can be utilized to identify gene expression changes that result in phenotypic alteration in plants (Kawakatsu et al., 2016). 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.

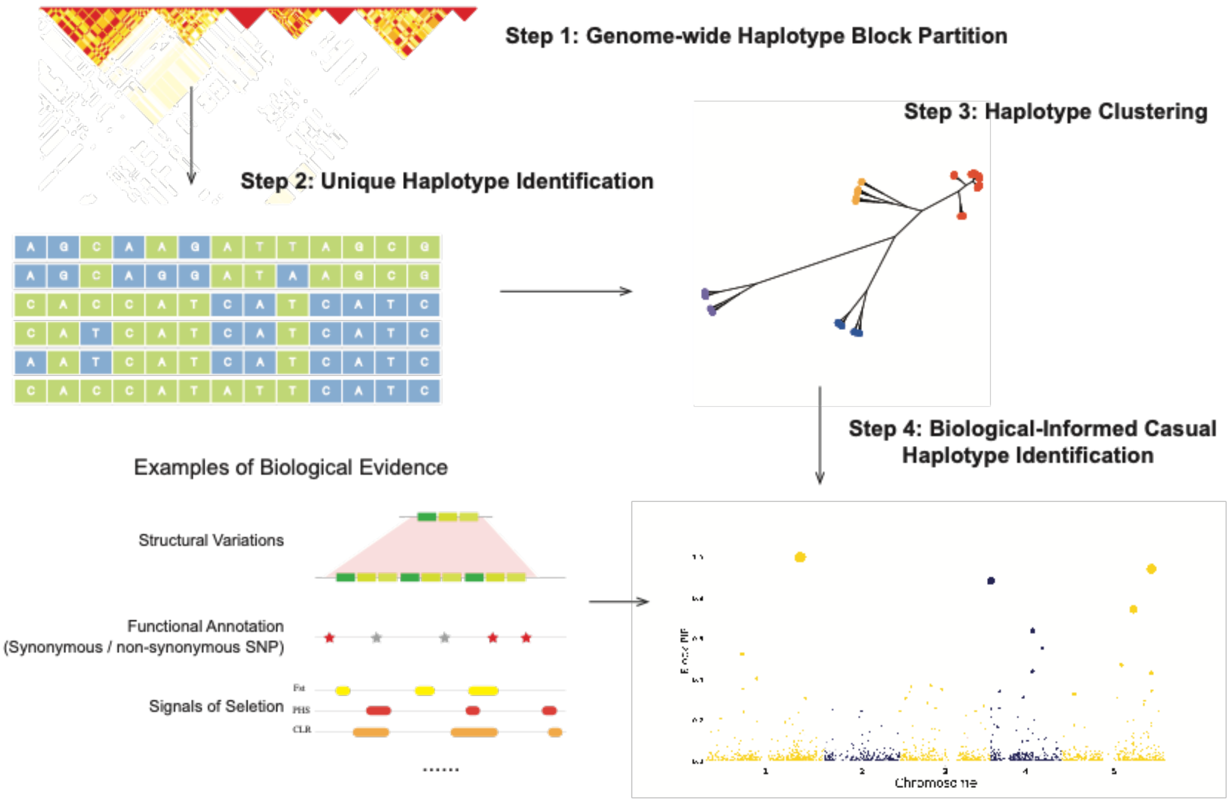
In this paper, we present a novel haplotype-based trait fine mapping framework, HapFM, that addresses limitations in plant GWAS methodologies. Unlike previous haplotype-based mapping algorithms, HapFM incorporates the use of unique haplotypes clusters based on historical recombination, rather than individual 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.

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 2.1). 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.


Step 1: Genome-wide Haplotype Block Partition
Step 2: Unique Haplotype Identification
Step 3: Haplotype Clustering
Step 4: Biological-Informed Casual Haplotype Identification
Examples of Biological Evidence
Structural Variations
Functional Annotation
(Synonymous / non-synonymous SNP)
Signals of Seletion
Fst
PHS
CLR
......
Block PIP
Chromosome
1
2
3
4
5

A	G	C	A	A	G	A	T	T	A	G	C	G
A	G	C	A	G	G	A	T	A	A	G	C	G
C	A	C	C	A	T	C	A	T	C	A	T	C
C	A	T	C	A	T	C	A	T	C	A	T	C
A	A	T	C	A	T	C	A	T	C	A	T	C
C	A	C	C	A	T	A	T	T	C	A	T	C

Figure 2.1. 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

Various algorithms were benchmarked to assess the robustness of block partitioning and haplotype clustering steps used in HapFM. Four clustering methods: affinity propagation (Frey and Dueck, 2007), X-means (Pelleg and Moore, 2000), KNN-spectral clustering and local-spectral clustering (Von Luxburg, 2007), 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 2.2a, Supplemental Figure 2.1a). 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 2.1b). Different clustering algorithms resulted in similar true positive rate in both low and high polygenicity simulations (Supplemental Figure 2.2). Affinity propagation gave 2.7 times more clusters than X-means in real data analyses, which costs longer computational time in the mapping step. 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 2.3). 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 2.2a). Minor mapping power differences were found between BigLD and Plink blocks in the low haplotype diversity simulations. BigLD blocks consistently resulted in higher or comparable mapping power than that of Plink blocks in all four QTL architectures in both low and high polygenicity simulations (Figure 2.2b, Supplemental Figure 2.4a). 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 2.2c, Supplemental Figure 2.4b). Plink blocks resulted in similar mapping power as that of uniform partitions.

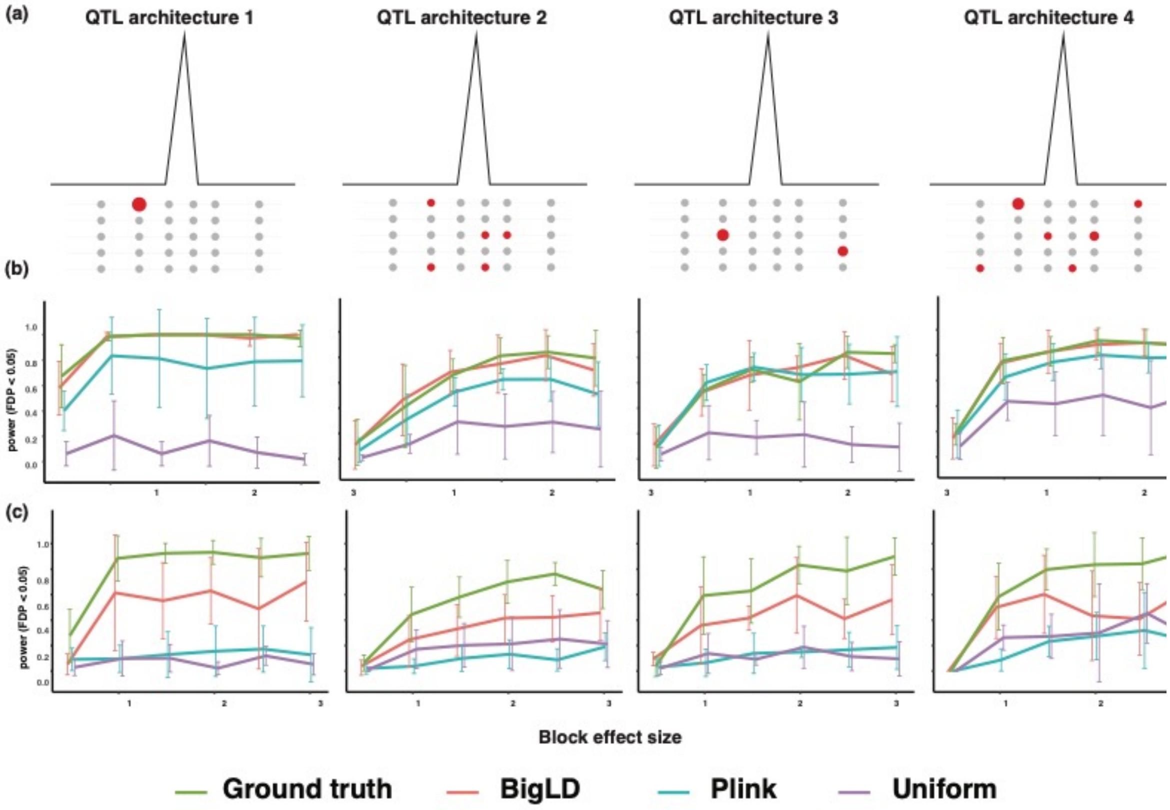


Figure 2.2. Simulation schemes and mapping power comparison of different block partition algorithms.

(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 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 low 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 low 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 2.5a). GEMMA, HapFM, and GMMAT resulted in similar true positive rates, which were significantly higher than that of BSLMM (Supplemental Figure 2.6a).

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 2.5b). 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 2.3). As 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 2.7).

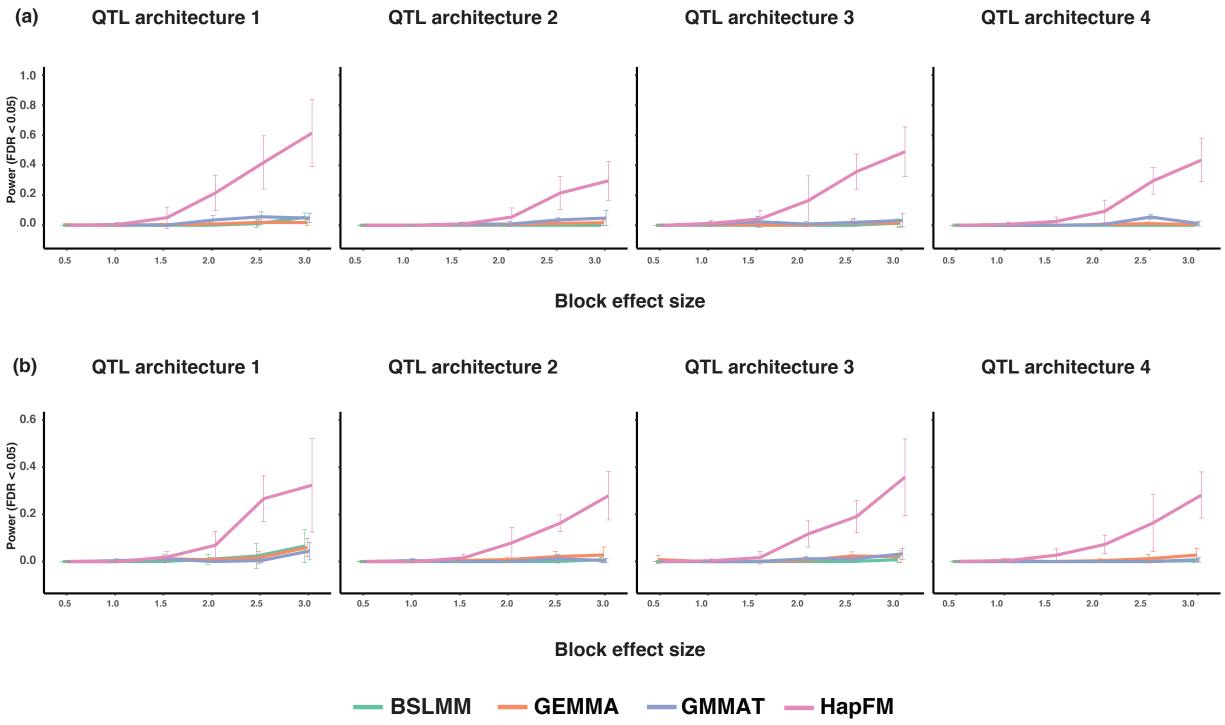


Figure 2.3. Mapping power comparisons of different GWAS algorithms in the high polygenicity simulations. The x-axis indicates the per-locus heritability.

(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 2.4a) in the high haplotype diversity simulation. When the trait polygenicity was high, the average interval length of GEMMA 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 2.4b). 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 2.1).

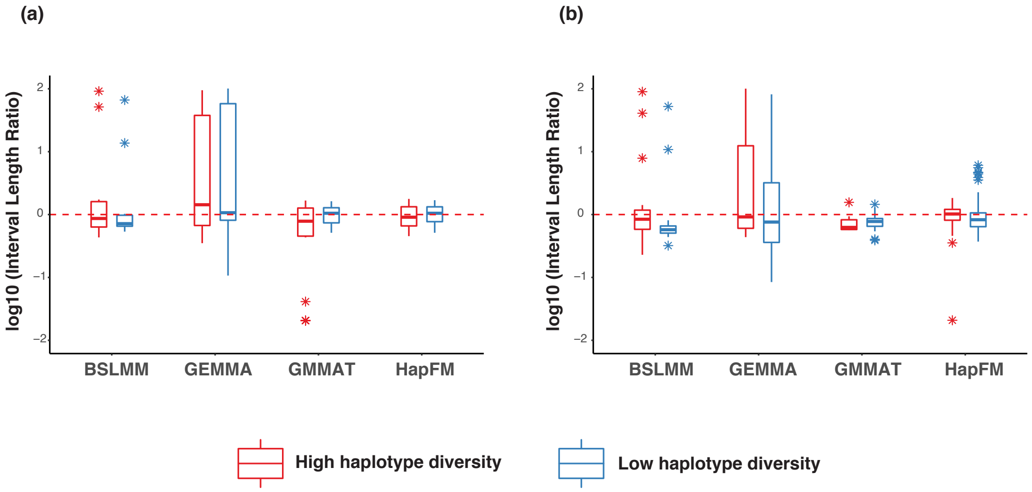


Figure 2.4. Mapping interval comparisons of different GWAS algorithms in the simulations. The interval length ratio was calculated by normalizing to the average HapFM's interval length. The red dash line indicates the average interval length of significant signals identified by HapFM.

(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

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 2.1). HapFM identified the most significant loci compared to the other GWAS algorithms in the Arabidopsis flowering time (FT10) dataset (Figure 2.5). 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 (Wang et al., 2008). 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 (Jakoby et al., 2008) (Pagnussat et al., 2005). 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 (Klepikova et al., 2016). 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 (Wang et al., 2009).

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 significant locus, FT locus, on Chromosome 1 (Figure 2.5). 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 2.1). 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 2.1). HapFM consistently resulted in similar or higher number of significant loci compared to 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 2.6a). HapFM-anno identified nine significant loci in total using biological-informed priors (Figure 2.6b,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 2.6b). The interval Chr1@7884994-7886542 is at the upstream region of gene AT1G22330 that is highly expressed in mature flowers (Klepikova et al., 2016). The interval Chr1@11474330-11475120 is at the upstream of the gene AT1G31940 that is highly expressed in mature flowers (Klepikova et al., 2016) and involved in seed germination (Narsai et al., 2011). 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 (Klepikova et al., 2016). The locus Chr5@23204856-23205070 overlaps with the gene AT5G57280 that has been shown to be highly expressed in different flower tissues (Klepikova et al., 2016) and pre-meristematic cell-mound formation during shoot regeneration (Shinohara et al., 2014). Two HapFM identified loci: Chr5@14063228-14197451 and Chr5@16141604-16146257, were not significant after incorporating biological annotations.

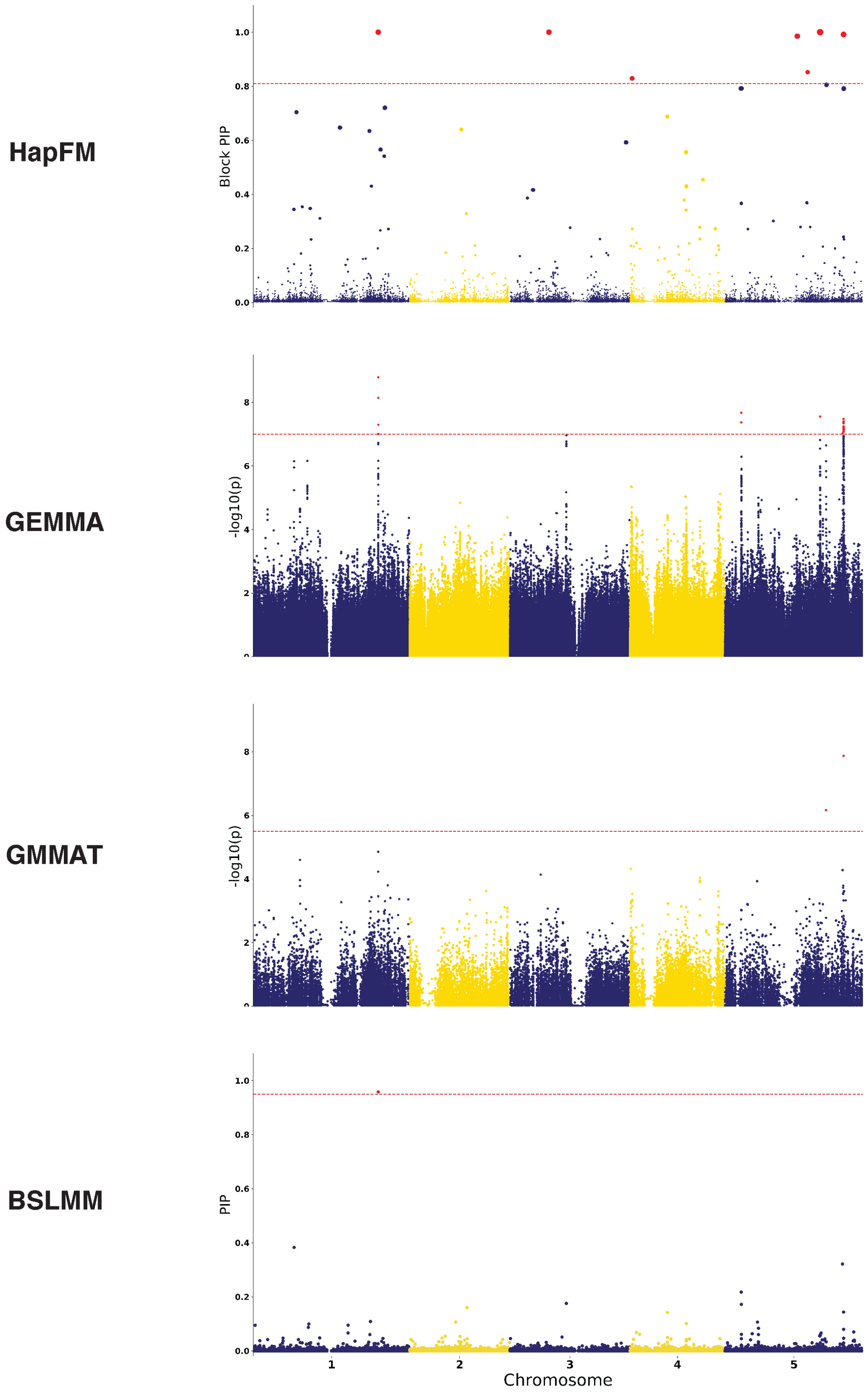

HapFM
Block PIP
1.0
0.8
0.6
0.4
0.2
0.0
GEMMA
-log10(p)
8
6
4
2
0
GMMAT
-log10(p)
8
6
4
2
0
BSLMM
PIP
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
Chromosome

Figure 2.5. 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	4	237,772	105
			BSLMM	1	21,863	80
			GMMAT	2	10,110	27
Rice Heading Time	529 individuals 1.43M SNPs	14,301	HapFM	10	166,031	792
			GEMMA	1	1,981,206	2473
			BSLMM	3	1,753,229	1868
			GMMAT	5	414,951	5908
Cassava HCN	1134 individuals 24.75K SNPs	9,112	HapFM	3	62,018	7
			GEMMA	4	1,068,992	20
			BSLMM	0	NA	NA
			GMMAT	2	1,166,404	46
Maize Height	263 individuals 23.09M SNPs	98,723	HapFM	10	398,161	5068
			GEMMA	0	NA	NA
			BSLMM	0	NA	NA
			GMMAT	NA	NA	NA

Table 2.1. Summary of GWAS results on the five real plant datasets.

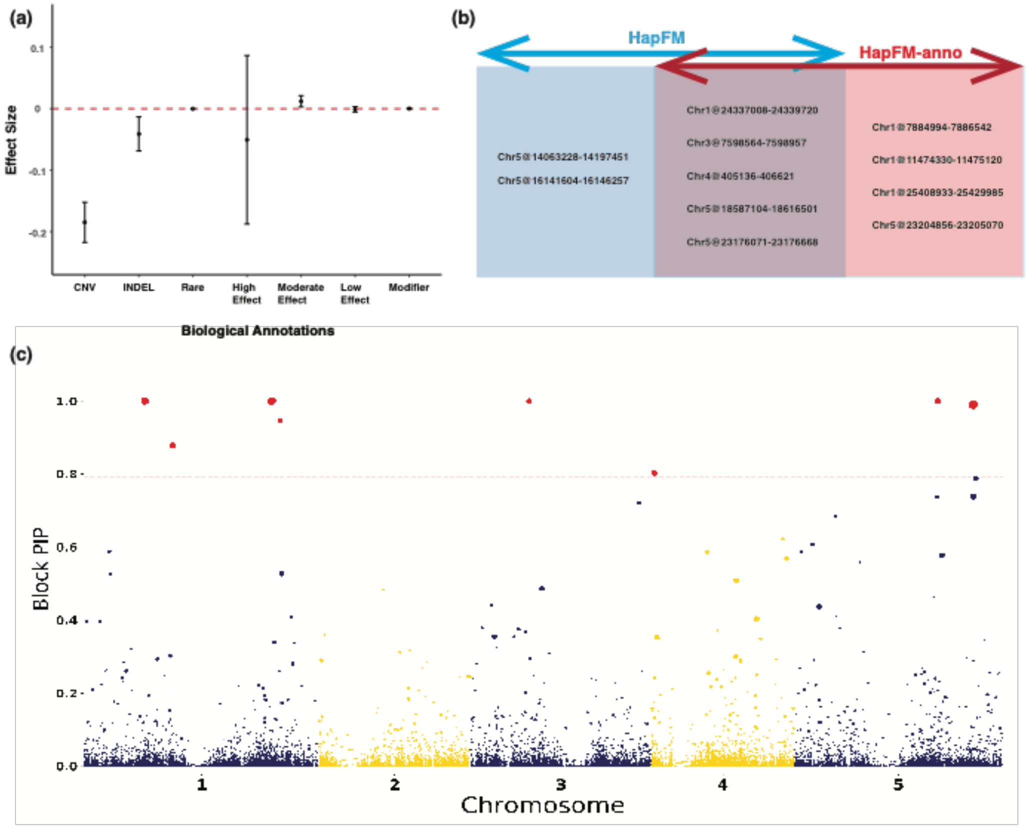


Figure2.6. Arabidopsis flowering time GWAS results using biological-informed priors (HapFM-anno). (a) The estimated effect sizes of different biological annotations for the Arabidopsis flowering time dataset. (b) The comparison of significant loci identified with and without incorporating biological 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.

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 (Cortes et al., 2021; Korte and Farlow, 2013). 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 (Tran et al., 2019), plant architecture (Yano et al., 2019), metabolic pathways (Tieman et al., 2017). 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(Wei et al., 2017). 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 (Chan et al., 2015).

Aggregating SNP effects is another way of increasing mapping power, such as SNP-set based method. This assumes that there may exist more 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 (Wu et al., 2019). 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 (Wang et al., 2017) (Ding and Kullo, 2007). 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 (Doebley et al., 2006) 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 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 (Cortes et al., 2021; Schaid et al., 2018; Zhou and Huang, 2019). 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 (Jaganathan et al., 2020). 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 (Westra et al., 2018) (Ferreiro-Iglesias et al., 2018). 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 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 (Liu et al., 2019), 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 2). 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 (Hamazaki and Iwata, 2020). 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 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 (Lei et al., 2021), 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 haplotype diversity, also known as haplotype blocks (Gabriel et al., 2002). 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 (Purcell et al., 2007) and BigLD (Kim et al., 2018). 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 (Novikov, 2019). 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:

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

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

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

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

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

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

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

As shown in the model, the haplotype effect sizes follow a mixture of normal density with mean 0 and variance σ_1^2 and a normal density with variance σ_0^2 pre-specified close to 0. The latent variable γ 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 β indicates the inferred probability of the haplotype block being causal.

The parameter π 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:

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

where $\Phi^{-1}$is the inverse of cumulative distribution function of a standard normal distribution, A is the matrix containing the annotation features, and θ is the vector of effect size corresponding to each biological annotation. The inference of θ follows the data augmentation method from (Albert and Chib, 1993).

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:

$$f_s = Zf_h$$

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:

$$y = C\alpha + X\eta + \epsilon,$$

whereby the coefficients α were sampled from a Uniform (-1, 1) distribution, and the entries in the covariate matrix C were sampled from a Uniform (-5, 5) distribution. X represents the simulated SNP genotype matrix. η represents the SNP effect sizes which was simulated in a hierarchical manner: causal blocks and causal SNPs in the block. At the block level, the probability, π_B, of a block containing true causal SNPs was simulated at 0.005 and 0.05. and the block effect size η_{B_j} were simulated ranging from 0.5 to 3. Under each true causal block, four types of architectures of true causal SNPs $(\lambda_i = 1)$ were simulated (Figure 2.1a):

(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, η_i, was assigned to each individual causal SNP based on the equation $\beta_{B_j} = \sum_{SNP_i \in B_j} \beta_i \, \mathrm{I}(\lambda_i = 1)$, where 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) (Seren et al., 2017), tomato metabolite (Zhu et al., 2018), rice yield (Xie et al., 2015), maize height (Peiffer et al., 2014) and a cassava HCN content (Ogbonna et al., 2021). The Arabidopsis flowering time GWAS dataset included genotype information from two previously published datasets: Arabidopsis Regmap (Horton et al., 2012) and 1001

Arabidopsis genome (Genomes Consortium. Electronic address and Genomes, 2016). 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 (Browning and Browning, 2007) to impute missing data and phased genotypes by following a 2-step imputation procedure (Wu et al., 2019). 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 pruning was performed on the filtered genotypes using PLINK with parameter set as --indep-pairwise 1000 100 0.1 (Borile et al.). 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 (Zhu et al., 2018). The genotype data of the 441 tomato accessions were processed according to Wu et al. published workflow (Wu et al., 2019). 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/) (Zhao et al., 2015). 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 (Ogbonna et al., 2021). 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 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 (Zhou and Stephens, 2012), Bayesian Sparse LMM BSLMM v0.98.1 (Zhou et al., 2013), and SNP-set based association method SMMAT v1.3.1 (Chen et al., 2019a). 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 (Brzyski et al., 2017):

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

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 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.

Chapter 3. A haplotype-based algorithm for pan-genome cohort selection and evaluation

Abstract

As sequencing cost continues to decrease, an increasing number of studies have shown that a single reference genome is often inadequate in many in genomic analyses resulting in biased or inaccurate results. A pan-genome reference that combines the sequence information of multiple individuals can mitigate the bias caused by a single reference. The premises of building a pan-genome reference is to more comprehensively representing the diversity of the species. Importantly, pan-genome cohort selection is a major factor in the level of diversity representativeness. A novel haplotype-based pan-genome cohort selection algorithm is presented that uses haplotype information to guide cohort selection to maximize local diversity. The workflow of HapPS consists of genome-wide block partition, representative haplotype cluster identification, and cohort selection by the Genetic Algorithm. The benchmark study between HapPS and a global-distance-based method showed that HapPS outperformed in five evaluation metrics including the average coverage of the high-diversity gene-overlapping blocks. The GO term enrichment analysis of the genes in the most diverse blocks showed significant enrichment of environmentally responding genes. HapPS also provides a quantitative evaluation of the pan-genome cohort selection. The evaluation metrics focuses on genome-wide and priority block coverages. In conclusion, HapPS is a robust and customizable algorithm that provides systemic solutions to select and evaluate diverse representatives a pan-genome cohort.

Introduction

A high-quality reference genome has become an essential tool for many biological studies in a multitude of bacterial, fungal, plant, and animal organisms. High-quality reference genomes provide the genomic sequence and location about the regions of interest, which are key to bioinformatic analyses, including read mapping and sequence annotations and molecular experiments such as gene cloning and genome editing. Reference genomes are also the foundation for population and evolution genomic research, such as studying human migration and admixture and understanding crop domestication.

Often, assembling a reference genome is the starting point for further investigations into the genomic aspect of the species. In the early 21st century, the completion of draft genomes for several model species, including human (Venter et al., 2001), drosophila (Adams et al., 2000), Arabidopsis (Arabidopsis Genome, 2000), rice (Goff et al., 2002; Yu et al., 2002), maize (Schnable et al., 2009) is a major milestone in genomics. These genomes were expensive and time-consuming to generate and sometimes involved multi-national collaborations primarily due to the relatively short read length of Sanger and Illumina sequencing technologies and inefficient short-read genome assembly algorithm (Metzker, 2010; Nagarajan and Pop, 2013; Pop and Salzberg, 2008; Treangen and Salzberg, 2011). Therefore, only one genome would be generated to represent the reference genome of that species. Arabidopsis, for example, was the Col-0 ecotype chosen as the reference genome because it had the most extensive genetic resources available at that time. The soybean cultivar Williams 82 was chosen for the reference genome because of its yield performance (Schmutz et al., 2010). Many other organisms followed suit such as cotton (Li et al., 2015), octopus (Albertin et al., 2015) and zebrafish (Howe et al., 2013).

Using a single reference poses significant limitations to study population diversity. Sequence polymorphisms alone can account for millions of sequence differences between even closely related individuals (Genomes Consortium. Electronic address and Genomes, 2016; Lin et al., 2014; Lv et al., 2020a). Rare variations found in the population can be even more difficult to identify since the inherent inaccuracy of short-read sequencing with the accuracy of polymorphism discovery decreases significantly as the genetic distance to the reference genotype increases (Wu et al., 2019).

Single reference genome limitations are even more prominent in plant studies due to extensive structural complexity and the repetitive nature of plant genomes. This can often make the identification of regions and genes of interest challenging. For example, Rhg-1 locus that is responsible for soybean-cyst nematode susceptibility is present as a cluster of four genes in the reference Willams2 genome. The resistant allele of Rhg-1 in the genotype, PI88788, has ten copies of the same genes, indicating a role for copy number variation (CNV) in determining the resistance to soybean cyst nematode (Cook et al., 2012). Hence, it would be difficult to determine the resistance mechanism without a de novo genomic assembly of this region subject to CNV because reads collapse from missing copies in the reference. Present-Absence variation (PAV) that is frequently encountered in many plant genomes (Huang et al., 2021; Lam et al., 2010; Wang et al., 2018b) can be even more problematic. PAV contributes to significant read alignment inaccuracy caused by misalignment to other homologous regions in the genome.

Multiple reference genomes for a species can provide robust solutions to many limitations associated with a single reference genome. As the accuracy and affordability of long-read sequencing technology increase, de novo reference-quality assemblies can be done at a reasonable cost. Accordingly, an increasing number of high-quality assemblies of non-reference genotypes have been published in recent years (Choi et al., 2020; Du et al., 2017; Hufford et al., 2021; Kim et al., 2021; Lin et al., 2021; Valliyodan et al., 2019). Their utility has underscored the importance of having more multiple reference genomes. This concept of multiple assemblies, referred to as pan-genomes, incorporates genome information from multiple individuals to mitigate the issues associated with single reference bias. Many plant pan-genomes have been constructed to facilitate population and functional genomics analyses (Della Coletta et al., 2021; Lei et al., 2021). One example is a recent soybean pan-genome assembled from 26 de novo genome assemblies using PacBio long-reads (Liu et al., 2020). This pan-genome encapsulates information about large structural variations and gene fusion events, which will help associate complex structural variations to agronomic traits.

Building a pan-genome reference consists of a series of optimization problems, starting from pan-genome cohort selection to choosing a suitable data structure for the pan-genome reference. This paper focuses on the problem of cohort selection. The conventional selection algorithms rely on global distance to select one individual for each group. In this paper, we proposed an alternative approach aiming to maximize haplotype representativeness in the pan-genome cohort. The algorithm, HapPS, focuses on representing haplotype diversity in haplotype blocks, especially regions of high diversity and functionally importance. In comparative studies, HapPS significantly outperformed the global-distance-based method in representing high-diversity regions and resulted in higher average genome-wide coverage. The results showed that HapPS provided better quantitative representativeness of the pan-genome cohort than global-distance-based methods. Finally, HapPS can be employed sequentially to update existing pan-genomes by selecting additional individuals with representative haplotypes previously absent. Overall, HapPS provides an alternative view of pan-genome references by prioritizing regions of interest and maximizing haplotype diversity in the cohort.

Results

Overview of HapPS workflow

We present a novel haplotype-based pan-genome cohort selection algorithm, HapPS, to serve as an alternative pan-genome cohort selection algorithm to global-distance-based methods. There are mainly three steps in the HapPS workflow: genome-wide haplotype block partition, block-wise representative haplotype identification, and pan-genome cohort selection by Genetic Algorithm (Figure 3.1). HapPS takes genotype information of a population as data input and then partitions the linear genome into non-overlapping segments of SNPs that reflect adjacent LD structures among SNPs. The default partitioning parameters are designed to reduce the number of single-SNP blocks. Users can adjust the size of haplotype blocks by choosing different r2 cutoffs. Each haplotype block is then treated as an independent unit for the subsequent analyses. Next, representative haplotypes of each block are identified by clustering unique haplotypes in the block. This step aims to avoid selecting genetically similar haplotypes more than once and maximize the represented diversity in each block. Each haplotype cluster is considered a representative. The individual and representative haplotype relationship is encoded into a haplotype design matrix containing the count information of each representative haplotype. Finally, HapPS separates genome-wide blocks into priority and secondary blocks according to the user's input and then uses Genetic Algorithm to select the pan-genome cohort prioritizing maximizing haplotype representativeness of priority blocks. The overall representativeness is quantified using a weighted sum of different metrics (see Material and Methods section). Once the Genetic Algorithm has finished, HapPS outputs a list of selected individuals for the pan-genome cohort and the selection evaluations.

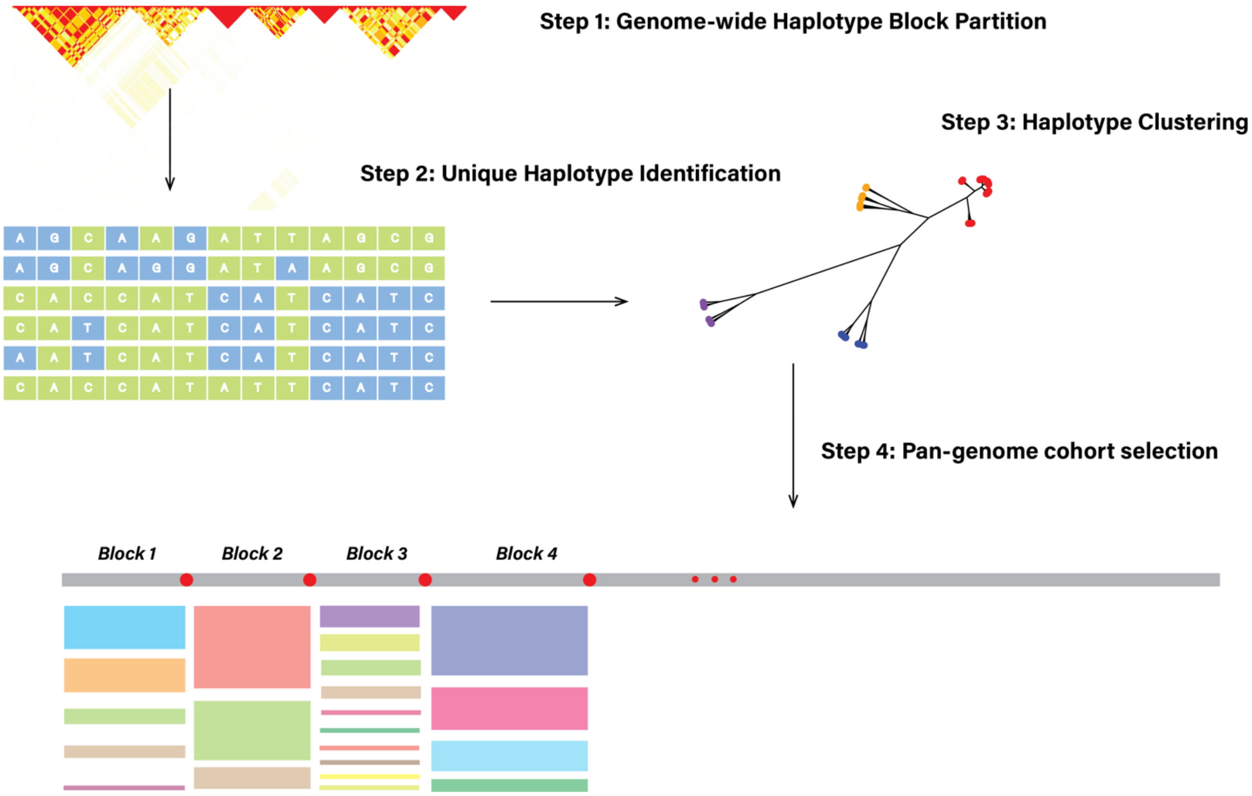

Step 1: Genome-wide Haplotype Block Partition
Step 2: Unique Haplotype Identification
Step 3: Haplotype Clustering
Step 4: Pan-genome cohort selection
A G C A A G A T T A G C G
A G C A G G A T A A G C G
C A C C A T C A T C A T C
C A T C A T C A T C A T C
A A T C A T C A T C A T C
C A C C A T A T T C A T C
Block 1
Block 2
Block 3
Block 4

Figure 3.1. The workflow of haplotype-based trait fine mapping (HapPS). HapFM consists of three steps: genome-wide haplotype block partition, unique haplotype clustering, and cohort selection by Genetic Algorithm

Proof-of-concept using Arabidopsis data

We performed a proof-of-concept study using an Arabidopsis dataset with 2021 Arabidopsis individuals and 170,977 bi-allelic SNP markers. The HapPS identified 19,130 haplotype blocks in total, 8049 of which were treated as priority blocks because they overlapped with at least one gene. The average block size was 7640 nt, the size of which increased in the centromeric regions because of high LD in this region of the genome (Figure 3.2).

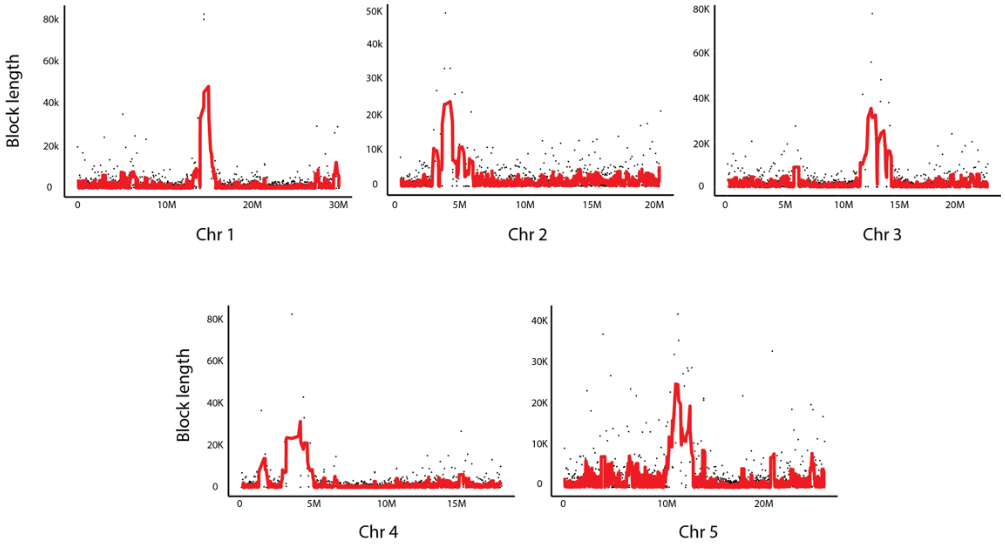


Figure 3.2. Block length distribution across the genome

The average haplotype clusters found in the priority blocks was 5.14 (±2.44 standard deviation), and the average haplotype clusters found in the secondary blocks was 2.44 (± 0.93 standard deviation). The diversity of the i^{th} block was defined as

$$Div_i = c \times \frac{\Sigma_{H_i} h^2}{\left(\frac{\Sigma_{H_i} h}{c}\right)^2 \times c}$$

Where c is the number of haplotype clusters in the i^{th} block, H_i is the i^{th} haplotype cluster in the block, h is the count of unique haplotypes in H_i. Significant differences were found between the block diversity of priority blocks and secondary blocks (Figure 3.3, Kruskal-Wallis rank sum test, *p*-value < 2.2e-16).

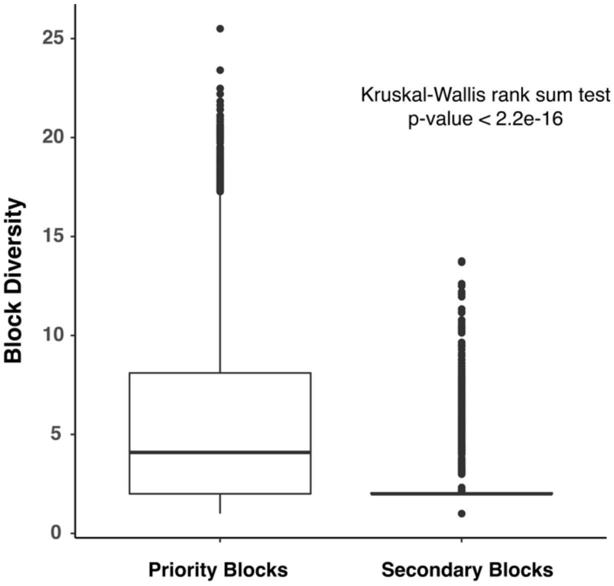


Figure 3.3. Haplotype cluster comparison between priority and secondary blocks.

A total of 6400 genes overlapped with the high-diversity priority blocks (top 5% quantile of priority block diversity). In total, 97 GO terms were found to be significantly enriched in the high-diversity priority blocks (Supplemental Table 3.1), 26 of which were enriched for at least 1.5 folds (Table 3.1). Bonferroni correction was used to adjust for the multiple comparisons.

GO biological process complete	Fold Enrichment	Adjusted *p*-value
cellular polysaccharide metabolic process	1.88	8.72E-03
cellular carbohydrate metabolic process	1.76	3.77E-03
polysaccharide metabolic process	1.71	4.85E-03
cellular amino acid metabolic process	1.76	2.64E-03
carboxylic acid metabolic process	1.59	1.75E-05
oxoacid metabolic process	1.54	2.64E-05
organic acid metabolic process	1.56	5.82E-06
small molecule metabolic process	1.57	1.43E-10
cell cycle process	1.75	2.10E-03
cell cycle	1.75	6.47E-04
response to light stimulus	1.73	3.15E-07
response to radiation	1.7	1.10E-06
response to abiotic stimulus	1.5	9.44E-12
carboxylic acid biosynthetic process	1.67	1.51E-02
organic acid biosynthetic process	1.69	3.52E-03
small molecule biosynthetic process	1.83	1.25E-07
cell wall organization or biogenesis	1.64	4.02E-03
carbohydrate derivative metabolic process	1.6	1.12E-04
embryo development ending in seed dormancy	1.59	1.26E-02
seed development	1.55	1.43E-03
fruit development	1.54	1.35E-03
post-embryonic development	1.54	1.89E-09
embryo development	1.57	1.56E-02
response to temperature stimulus	1.54	1.59E-02
organonitrogen compound biosynthetic process	1.51	4.06E-07
shoot system development	1.5	1.39E-02

Table 3.1. GO terms with at least a 1.5-fold of enrichment in the genes in high-diversity priority blocks

Benchmark HapPS against the global-distance algorithm

Benchmark analyses of HapPS and a global-distance-based selection algorithm were performed using the Arabidopsis dataset. There were five evaluating metrics: the average coverage of genome-wide blocks, the percentage of genome-wide blocks with coverage greater than 80%, the average coverage of priority blocks, the percentage of priority blocks with coverage greater than 80%, and the average coverage of high-diversity priority blocks (top 5% quantile of priority block diversity). HapPS was repeated 30 times for each selection, and average values were used for the comparison analyses due to the heuristic nature of the Genetic Algorithm. The comparison results showed that HapPS consistently outperformed the conventional global-distance-based selection algorithm in all five metrics (Figure 3.4). The largest difference between the two algorithms was found in the coverage of the high-diversity priority blocks. In these high-diversity priority blocks, HapPS, on average, was able to result in 1.51 (standard deviation 0.26) times coverage of that of the global-distance-based algorithm. Moreover, HapPS was able to reach similar evaluation metrics by selecting fewer individuals. For example, HapPS needed to select ten individuals, whereas a global-distance-based algorithm needed 20 individuals to achieve the average 70% coverage of the high-diversity priority blocks.

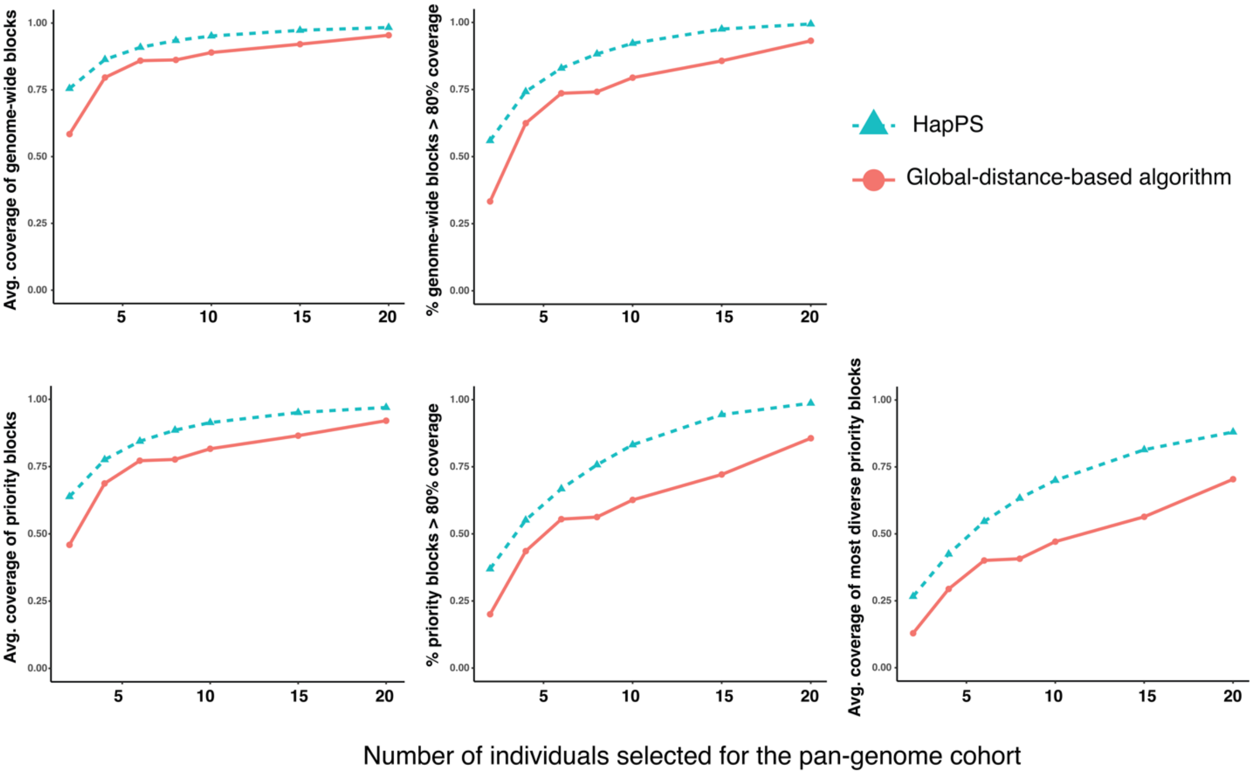


Figure 3.4. Comparison of HapPS and a global-distance-based algorithm on five different metrics.

Conclusions and Discussions

One primary reason for building a pan-genome reference is to mitigate the bias associated with single reference genomes. Pan-genome consists of a core part of the genome containing sequences shared between most individuals of the species and a dispensable part of the genome containing regions found in only a subset of the population. The genes in the core genome are often highly conserved genes with essential functions. In contrast, the genes in the dispensable genome are often found to be enriched for environmental responsive genes (Bayer et al., 2020; Della Coletta et al., 2021; Hufford et al., 2021; Lei et al., 2021; Liu et al., 2020). Therefore, it would be beneficial to represent these more variable sequences to maximize the pan-genome reference's representativeness. The presence-absence and copy number variations can also introduce significant biases when analyzing non-reference genotypes. The gene enrichment results of the high-diversity gene-overlapping regions suggested that this bias may be significant for the genetic studies of environmental adaptation. For that reason, it is essential that the pan-genome reference should maximize representation of the highly diverse regions of the genome. A pan-genome reference should balance the complexity of the pan-genome and the representativeness of the species.

Cohort selection is the first step to building a pan-genome reference, as this determines the level of representativeness. HapPS represents a novel haplotype-based pan-genome cohort selection algorithm that maximizes the local haplotypes diversity in the cohort. Unlike the conventional global-distance-based selection algorithms, HapPS partitions the linear genome into non-overlapping blocks and prioritizes the cohort selection providing better coverage of regions of high interest. To the best of our knowledge, this is the first algorithmic and systemic approach designed for selecting a pan-genome cohort.

HapPS has other advantages over global-distance-based selection methods besides maximizing representativeness. First, the algorithm can be customized to set parameters optimized for different species and research interests. These parameters include the LD cutoff for block partitions, the list of priority blocks, and the weights for selection criteria. Another advantage of HapPS is prioritizing regions of interest, which is difficult to achieve by global-distance-based algorithms. In any pan-genome cohort, the number of individuals is limited by resources. Hence, the pan-genome cohort is unlikely to represent the whole genome complexity found in a species. HapPS was designed to maximize the diversity coverage of priority regions with a trade-off on the coverage of secondary regions. In the proof-of-concept study, HapPS was shown to provide greater coverage to high-diversity blocks than the global-distance-based method. Accordingly, these blocks were shown to be significantly enriched for genes associated with environmental GO terms (Table 1). These results indicated that HapPS provided higher coverage of the genetic diversity in these regions than the conventional global-distance-based methods.

In summary, HapPS provides a customizable method of pan-genome cohort selection that maximizes local haplotype representativeness for the cohort. One future improvement will be incorporating complex structural variations into the cohort selection process. The inclusion of structural variations will likely increase the accuracy of haplotype clustering to represent genetic diversity within the species better.

Material and Methods

Details of the HapPS workflow

The HapPS workflow includes genome-wide haplotype block partition, unique haplotype clustering, and Genetic Algorithm selection of the cohort panel. The details of each step are described below.

Initially, genome-wide haplotype block partition takes place in the HapPS workflow to ouput a non-overlapping set of SNPs that represent the haplotype block structure in the genome. This is done using a 2-step algorithm to accelerate the block partition process. The first step partitions the whole genome into the large independent blocks by setting minimum pairwise SNP LD cutoff at $r^2 > 0.1$. The second step then partitions each independent block into correlated haplotype blocks using BigLD (Kim et al., 2018). HapPS changes the default parameters of BigLD to "CLQmode = maximal, CLQcut = 0.5" to avoid numerous single-SNP blocks. Users can also input their own haplotype block partitions.

In the next step, for each haplotype block identified in the previous step, a set of unique haplotypes are enumerated and clustered based on their genetic similarity. Manhattan distances between pairs of unique haplotypes are calculated. The affinity matrix is calculated next using the formular, $W_{ij} = e^{\frac{-d_{ij}^2}{2\sigma_i\sigma_j}}$, where W_{ij} is the affinity between haplotype i and haplotype j, d_{ij} is the Manhattan distance between haplotype i and haplotype j, σ_i is the standard deviation of Manhattan distance of haplotype i to all the other haplotypes. The K-nearest neighbor graph is then built based on the calculated affinity matrix. If not specified by the user, HapPS determines k using the formular $\max(5,\ c/50)$, where c is the total number of unique haplotypes. HapPS then applies Louvain's method (Blondel et al., 2008) to group similar haplotypes by maximizing the modularity of the clusters. Each cluster is then treated as a representative haplotype, and a design matrix is constructed to indicate the number of occurrences of each representative haplotype in each individual.

The last step in the HapPS workflow is performing cohort selection using Genetic Algorithm (Mitchell, 1996). Block coverage is defined as the sum of the frequency of existing representative haplotypes in the block of the selected cohort. The objective is to maximize the coverage of more essential blocks if maximizing every block is not feasible given the number of individuals selected. Therefore, HapPS requires the user to input a list of priority blocks, for example, blocks that overlap with genes, to give greater weight to the selection toward these regions. The rest of the blocks are then treated as secondary blocks. HapPS uses six metrics to guide the selection, and the objective function for the Genetic Algorithm is to maximize the weighted sum of all the metrics. The metrics and weights are described in Table 2, and users can customize the weights. Simple Genetic Algorithm is used for cohort selection, and the initial population size, crossover rate, and mutation rate are set to 100, 0.2, 0.2 by default.

Metric	Weights
average coverage of top 5% most diverse priority blocks	20
percentage of all priority blocks that has over 80% coverage	10
average coverage of all the priority blocks	5
the average coverage of top 5% most diverse secondary blocks	10
percentage of all the secondary blocks that has over 80% coverage	5
average coverage of all secondary blocks	3

Table 3.2. The selection metric and weights for the Genetic Algorithm

Generating an ad hoc Arabidopsis population for pan-genome selection

Two Arabidopsis SNP datasets, the 1001 Arabidopsis and Regmap genotype datasets, were combined to generate an *ad hoc* population for pan-genome cohort selection. In the 1001 Arabidopsis genotype dataset, non-biallelic SNPs and SNPs with missing data percentage greater than 20% were omitted yielding a total of 8,231,757 remaining SNPs. In the Regmap genotype dataset, SNPs that were not in LD ($r^2 < 0.1$) with nearby 20 SNPs were filtered out leaving 202,339 remaining SNPs, 170,977 of which were also included in the filtered 1001 Arabidopsis genotype dataset. Next, the two filtered datasets were combined to generate a new dataset of 2021 individuals with 170,977 SNPs. The missing data in the combined dataset was imputed by Beagle with default parameters (Browning and Browning, 2016).

Benchmark different pan-genome cohort selection methods

HapPS was benchmarked against a global-distance-based cohort selection method using the combined Arabidopsis SNP dataset. In the global-distance-based cohort selection method, pairwise IBS distances were first calculated using PLINK 1.9 (Purcell et al., 2007). The individuals were then hierarchically clustered according to their IBS distance, and the tree was pruned to a certain height to result in *m* groups corresponding to the number of individuals in the pan-genome cohort. The algorithm then randomly selected one individual from each group. HapPS used the default settings. The priority blocks were defined as blocks that overlapped with Arabidopsis genes (version TAIR10.49). BEDtools (Purcell et al., 2007) was used to identify the gene-overlapping blocks. Each algorithm selected the same number of individuals [4, 6, 8, 10, 12, 16, 20] and repeated 30 times.

The performance of the two selection algorithms was evaluated using five metrics: the average coverage of genome-wide blocks, percentage of the genome-wide blocks with over 80% coverage, the average coverage of priority blocks, percentage of the priority blocks with over 80% coverage, and the average coverage of top 5% diverse priority blocks.

Annotations of genes in the top 5% diverse priority blocks

Gene ontology enrichment analysis was performed on the genes in the top 5% diverse priority blocks using PANTHER 16.0. Bonferroni correction was applied to calculate adjusted *p*-values.

Conclusion and Future Direction

Powerful and robust computational algorithms are keys to uncovering hidden diversity in wild and unrelated individuals for crop improvement. The bioinformatic pipeline can be categorized into reference genome construction, variant discovery, and trait mapping steps. Many algorithms have been developed and optimized for each task using human genomic studies. However, their efficacies are largely depreciated due to fundamental differences between human and plant genomic datasets. In this dissertation, I have presented three chapters of studies to optimize and develop novel computational algorithms for plant diversity discovery. The focuses of these chapters are variant discovery pipeline, plant trait mapping, and pan-genome cohort selection algorithms.

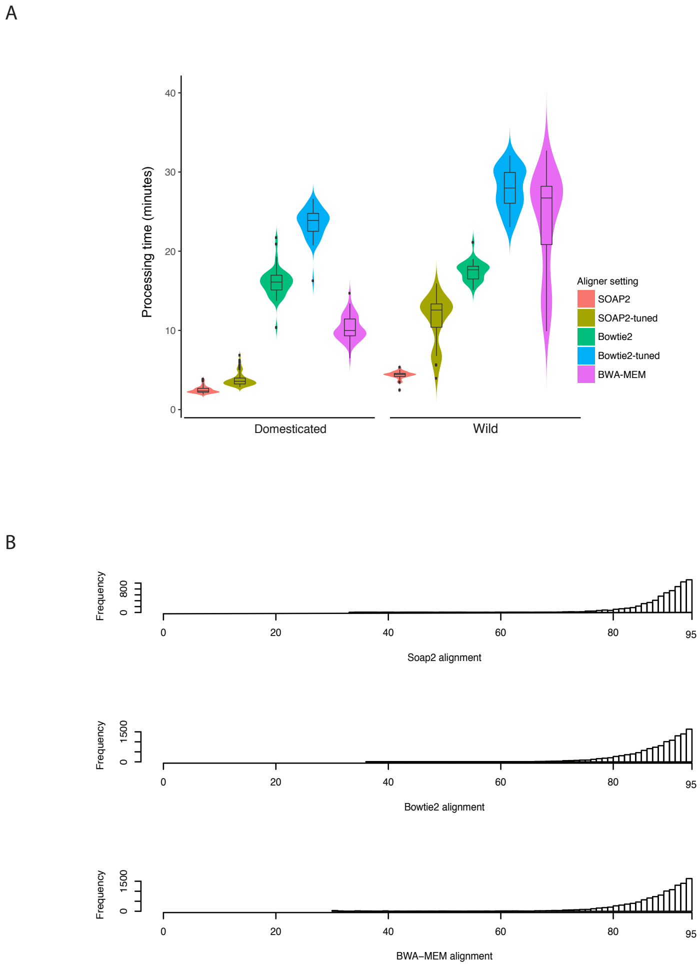
In Chapter 1, I performed a detailed evaluation of each step in the variant calling pipeline. I found that BWA-MEM was better at detecting more true-positive alignments, especially in distantly related samples, while Bowtie2 was better at minimizing the incorrect alignments. Incorporating multiple reference genomes gave a complete picture of variations, especially when the samples showed considerable presence-absence variations. For variant filtering, the optimal approach found in our test was to incorporate a combination of machine learning and hard filtering, in which a set of "known" SNPs was used as the training set for machine learning. This requires a panel of known, high-quality SNPs, however, which may be unavailable for many plant species. Finally, the importance of high-quality reference panels was emphasized during the imputation step especially when genotype imputation was challenging due to small LD blocks or not enough samples. Above all, the computational pipeline to discover variations from plant sequencing data will depend upon the diversity of the datasets, the depth of sequence coverage, and the availability of external resources such as reference panels and gold-standard SNPs.

In Chapter 2, I developed a novel haplotype-based trait fine-mapping algorithm, HapFM, to address specific trait mapping issues in plant studies. HapFM is a comprehensive and powerful mapping algorithm that includes genome-wide block partition, haplotype clustering, genome-wide statistical fine-mapping, and incorporating biological annotations. We demonstrated that HapFM resulted 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 the genetic architecture of traits.

In Chapter 3, I developed a haplotype-based pan-genome cohort selection algorithm, HapPS to maximize the haplotype representativeness in the pan-genome reference. Similar to HapFM, the workflow of HapPS includes genome-wide block partition, representative haplotype cluster identification, and cohort selection by the Genetic Algorithm. In contrast to the global-distance-based selection algorithm, HapPS uses local haplotype information to guide the selection. It prioritizes regions of interest to the purpose of the study and customizes parameters to fit the scope of the study. One of the goals for pan-genome is to capture environmental genes that are unique to a subset of the species. In the proof-of-concept study, the high-diversity priority regions are enriched in genes responding to environmental changes. HapPS resulted in significantly higher representativeness of these regions than the global-distance-based selection.

The future of plant genomics will transit into a pan-genomic era. A pan-genome reference will primarily benefit the diversity discovery and trait mapping in principle. One of the biggest obstacles is the availability of bioinformatic programs for pan-genomics. In my dissertation, I have shown the advantages of using haplotypes for mapping and pan-genome cohort selection. I want to explore the possibility of using haplotypes as units in pan-genomic studies. The pan-genomic pipeline starts from the haplotype-based pan-genome cohort selection. The block partition and representative haplotype can be used to construct the pan-genome graph. Next is to align reads to the closest haplotype and call haplotypes from the pan-genome. The haplotype information can be directly inputted into HapFM for trait mapping. I will be interested in realizing this haplotype-based pan-genome diversity discovery pipeline.

Supplemental Material

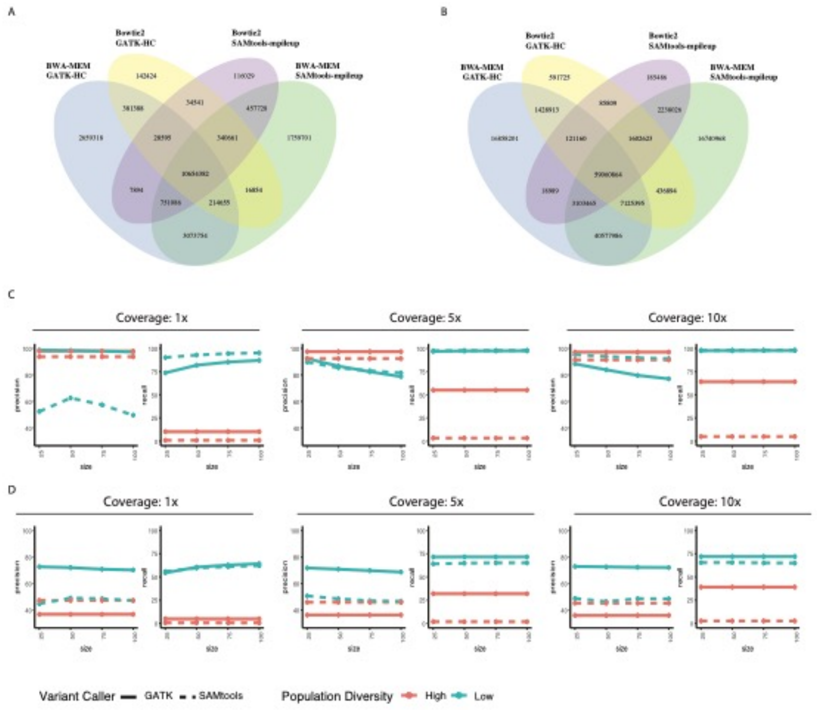


Supplemental Figure 1.1. Alignment time comparison of different aligners

(A) Alignment percentage of five different aligner settings: SOAP2, SOAP2-tuned, Bowtie2, Bowtie2-tuned and BWA-MEM calculated for domesticated tomatoes and wild relatives. The width of violin plot is proportional to the density of the data. Boxplots inside violin plot indicate quantiles and outliers.

(B) The alignment length distribution of different aligners. Only alignment shorter than 96 nt was plotted for better visualization.

Supplemental Figure 2



Supplemental Figure 1.2. Evaluation of variant calling programs using simulated plant genomic datasets

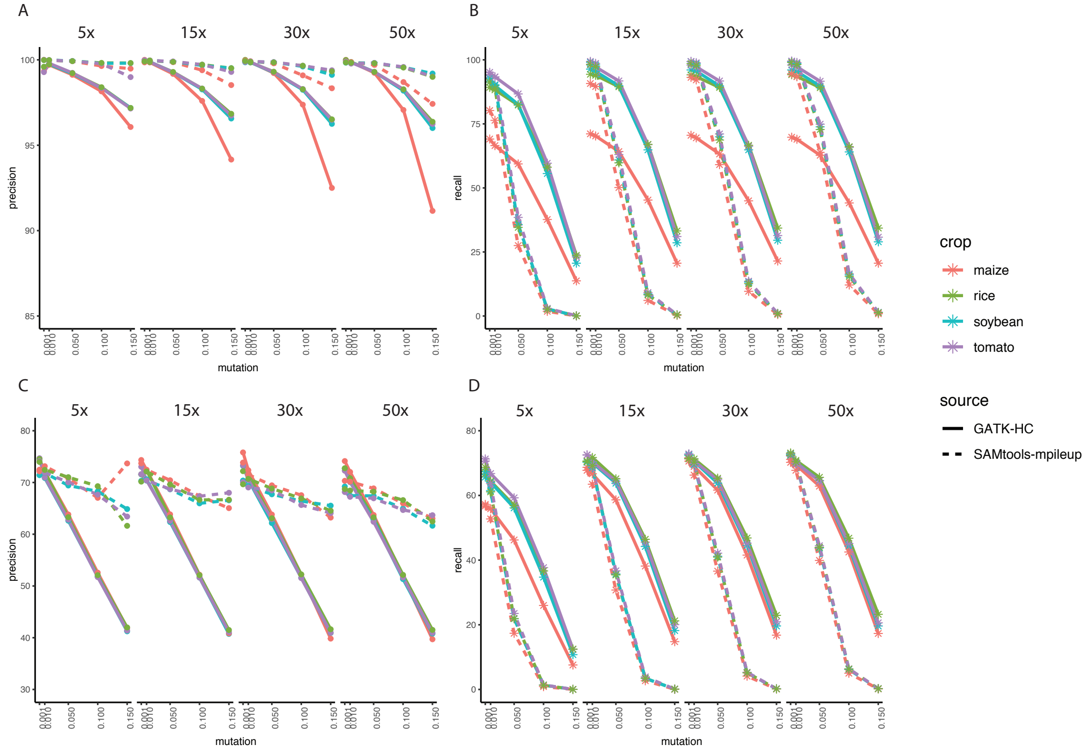
(C) The Venn diagram of SNPs identified from domesticated tomato dataset using different aligner and variant caller combinations.

(D) The Venn diagram of SNPs identified from wild tomato dataset using different aligner and variant caller combinations.

(E) The comparison of the performance of GATK-HC and SAMtools-mpileup on raw SNPs at different coverages, population diversity and population size.

(F) The comparison of the performance of GATK-HC and SAMtools-mpileup on raw INDELs at different coverages, population diversity and population size.

Supplemental Figure 3



Supplemental Figure 1.3. Evaluation of different variant calling programs on single genomic dataset

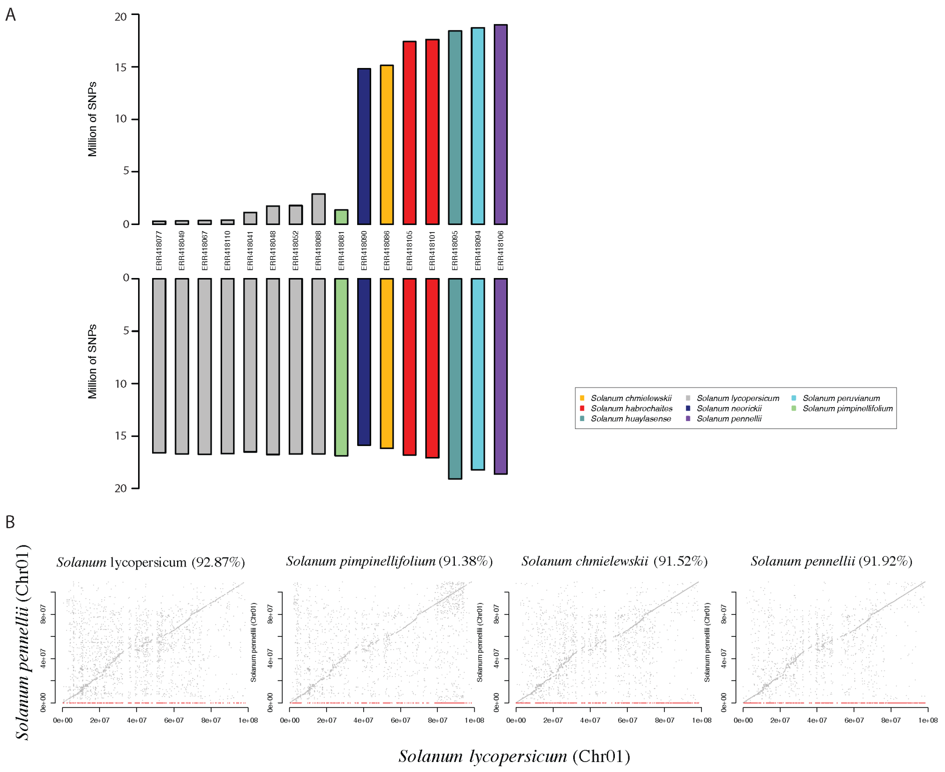
(A) SNP precision results of GATK-HC and SAMtools-mpileup on single simulated dataset with varied coverages, mutation rates and crop species

(B) SNP recall results of GATK-HC and SAMtools-mpileup on single simulated dataset with varied coverages, mutation rates and crop species

(C) INDEL precision results of GATK-HC and SAMtools- on single simulated dataset with varied coverages, mutation rate sand crop species

(D) INDEL recall results of GATK-HC and SAMtools-mpileup on single simulated dataset with varied coverages, mutation rates and crop species

Supplemental Figure 4

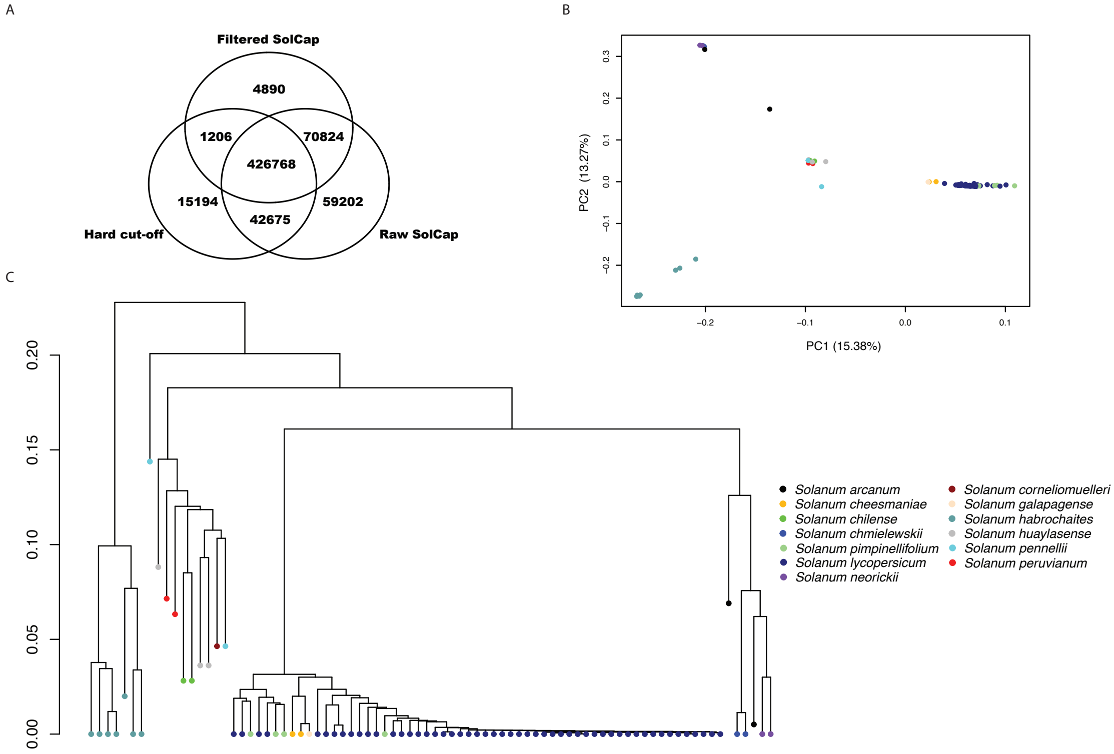


Supplemental Figure 1.4. Cross-reference comparison on SNP identification

(A) Number of SNPs identified using *S. lycopersicum* or *S. pennellii* genome assembly as the reference

(B) SNP identification of four tomato samples was performed in chromosome 1 in *S. lycopersicum* reference genome. The corresponding physical positions of SNPs in the *S. pennellii* reference was plotted. The grey dots represented the SNPs that were able to be located at the corresponding positions in *S. pennellii* genome, red dots represented the SNPs that were unable to be located to corresponding positions in *S. pennellii* genome. The percentage of corresponding SNPs are written next to the species name.

Supplemental Figure 5



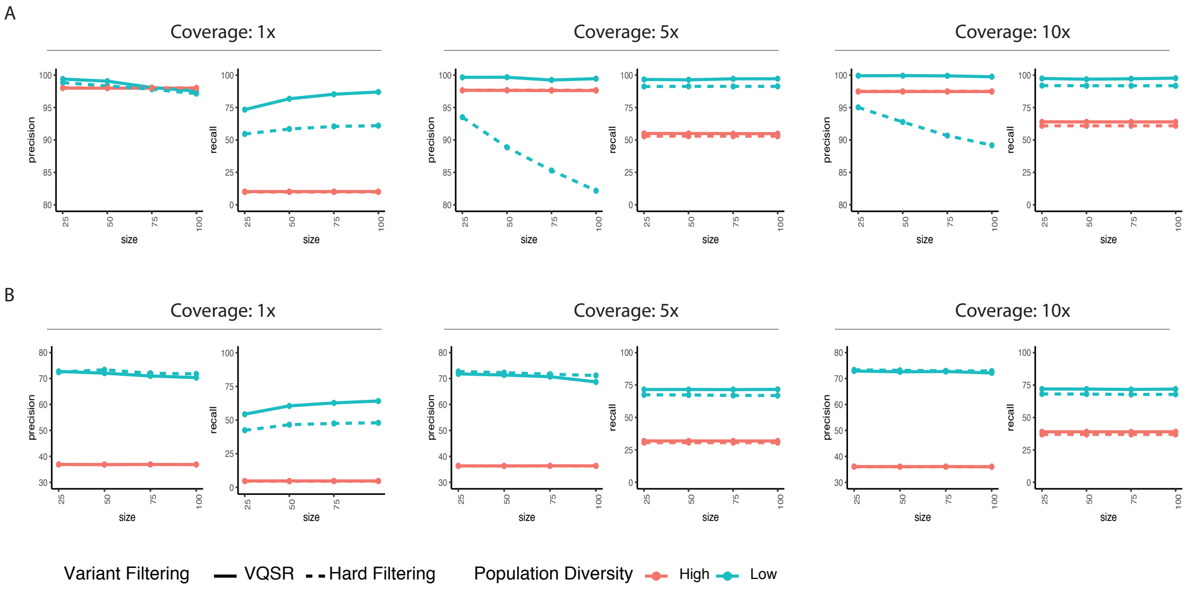
Supplemental Figure 1.5. Machine-learning based variant filtering

(A) Venn diagram of SNPs in the 10M region of Chromosome 1 using HARD, ML and COMBINED filtering methods

(B) Population structure of 82 tomato genomes using high-confidence SNPs

(C) IBS distance of 82 tomato genomes using high-confidence SNPs

Supplemental Figure 6

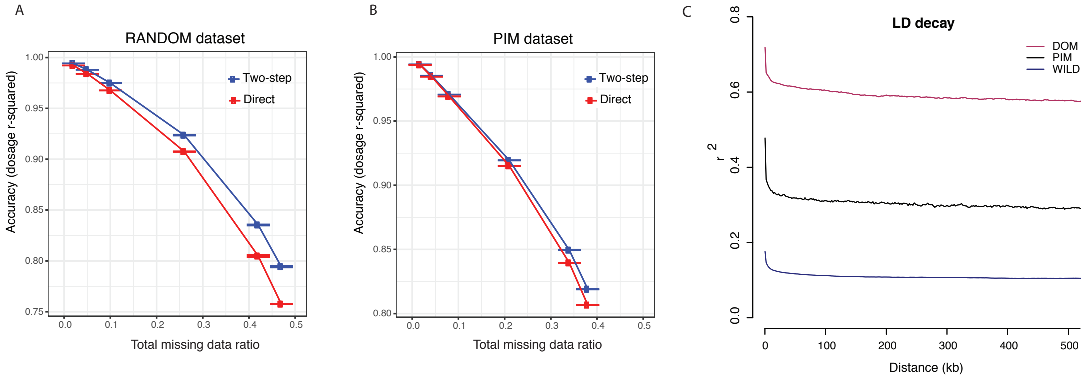


Supplemental Figure 1.6. Comparison between VQSR and hard-filtering

(A) The comparison of the performance of VQSR and hard-filtering on SNPs at different coverages, population diversity and population size.

(B) The comparison of the performance of VQSR and hard-filtering on INDELs at different coverages, population diversity and population size.

Supplemental Figure 7

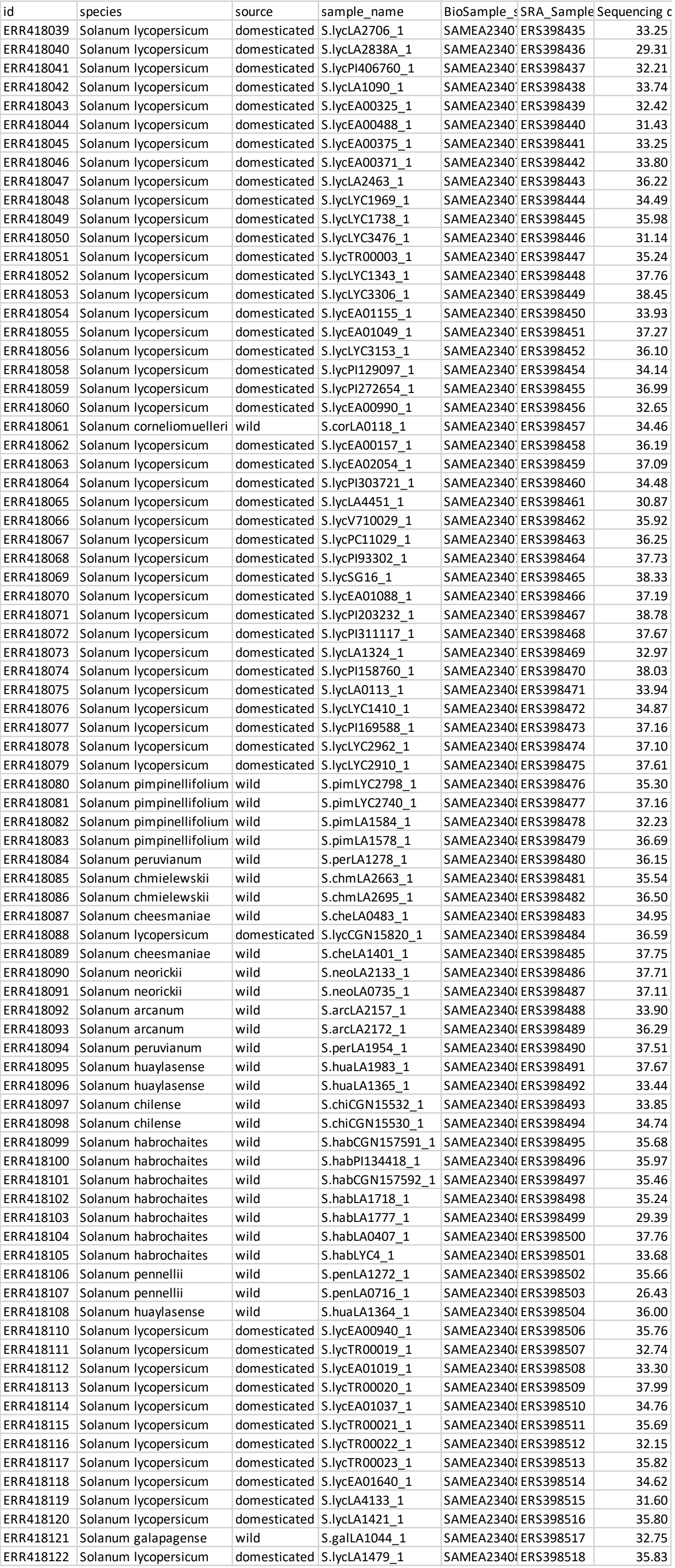


Supplemental Figure 1.7. Comparison between direct and two-step imputation

(A) Imputation accuracy using direct imputation and 2-step imputation relative to missing SNPs in 200 random tomato samples

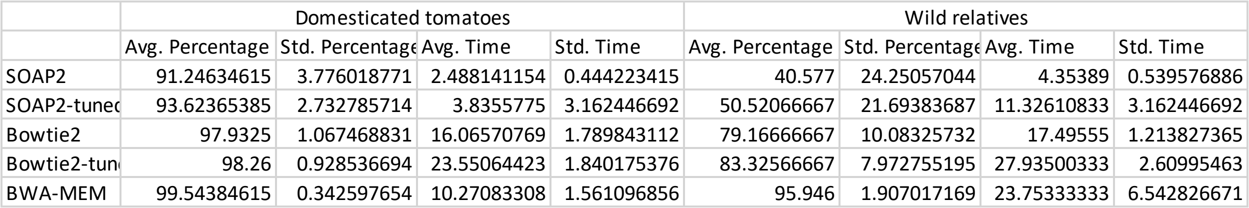
(B) Imputation accuracy using direct imputation and 2-step imputation relative to missing SNPs in 50 *S. pimpinellifolium* tomato samples

(C) Comparison of LD decay of SNPs from different populations.



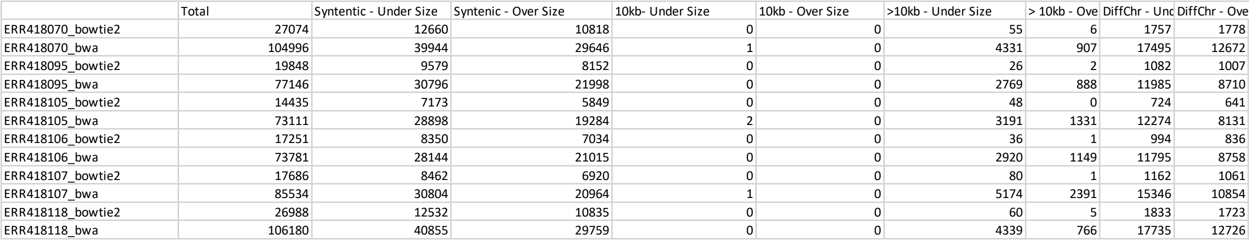
id	species	source	sample_name	BioSample_s	SRA_Sample	Sequencing c
ERR418039	Solanum lycopersicum	domesticated	S.lycLA2706_1	SAMEA23407	ERS398435	33.25
ERR418040	Solanum lycopersicum	domesticated	S.lycLA2838A_1	SAMEA23407	ERS398436	29.31
ERR418041	Solanum lycopersicum	domesticated	S.lycPI406760_1	SAMEA23407	ERS398437	32.21
ERR418042	Solanum lycopersicum	domesticated	S.lycLA1090_1	SAMEA23407	ERS398438	33.74
ERR418043	Solanum lycopersicum	domesticated	S.lycEA00325_1	SAMEA23407	ERS398439	32.42
ERR418044	Solanum lycopersicum	domesticated	S.lycEA00488_1	SAMEA23407	ERS398440	31.43
ERR418045	Solanum lycopersicum	domesticated	S.lycEA00375_1	SAMEA23407	ERS398441	33.25
ERR418046	Solanum lycopersicum	domesticated	S.lycEA00371_1	SAMEA23407	ERS398442	33.80
ERR418047	Solanum lycopersicum	domesticated	S.lycLA2463_1	SAMEA23407	ERS398443	36.22
ERR418048	Solanum lycopersicum	domesticated	S.lycLYC1969_1	SAMEA23407	ERS398444	34.49
ERR418049	Solanum lycopersicum	domesticated	S.lycLYC1738_1	SAMEA23407	ERS398445	35.98
ERR418050	Solanum lycopersicum	domesticated	S.lycLYC3476_1	SAMEA23407	ERS398446	31.14
ERR418051	Solanum lycopersicum	domesticated	S.lycTR00003_1	SAMEA23407	ERS398447	35.24
ERR418052	Solanum lycopersicum	domesticated	S.lycLYC1343_1	SAMEA23407	ERS398448	37.76
ERR418053	Solanum lycopersicum	domesticated	S.lycLYC3306_1	SAMEA23407	ERS398449	38.45
ERR418054	Solanum lycopersicum	domesticated	S.lycEA01155_1	SAMEA23407	ERS398450	33.93
ERR418055	Solanum lycopersicum	domesticated	S.lycEA01049_1	SAMEA23407	ERS398451	37.27
ERR418056	Solanum lycopersicum	domesticated	S.lycLYC3153_1	SAMEA23407	ERS398452	36.10
ERR418058	Solanum lycopersicum	domesticated	S.lycPI129097_1	SAMEA23407	ERS398454	34.14
ERR418059	Solanum lycopersicum	domesticated	S.lycPI272654_1	SAMEA23407	ERS398455	36.99
ERR418060	Solanum lycopersicum	domesticated	S.lycEA00990_1	SAMEA23407	ERS398456	32.65
ERR418061	Solanum corneliomuelleri	wild	S.corLA0118_1	SAMEA23407	ERS398457	34.46
ERR418062	Solanum lycopersicum	domesticated	S.lycEA00157_1	SAMEA23407	ERS398458	36.19
ERR418063	Solanum lycopersicum	domesticated	S.lycEA02054_1	SAMEA23407	ERS398459	37.09
ERR418064	Solanum lycopersicum	domesticated	S.lycPI303721_1	SAMEA23407	ERS398460	34.48
ERR418065	Solanum lycopersicum	domesticated	S.lycLA4451_1	SAMEA23407	ERS398461	30.87
ERR418066	Solanum lycopersicum	domesticated	S.lycV710029_1	SAMEA23407	ERS398462	35.92
ERR418067	Solanum lycopersicum	domesticated	S.lycPC11029_1	SAMEA23407	ERS398463	36.25
ERR418068	Solanum lycopersicum	domesticated	S.lycPI93302_1	SAMEA23407	ERS398464	37.73
ERR418069	Solanum lycopersicum	domesticated	S.lycSG16_1	SAMEA23407	ERS398465	38.33
ERR418070	Solanum lycopersicum	domesticated	S.lycEA01088_1	SAMEA23407	ERS398466	37.19
ERR418071	Solanum lycopersicum	domesticated	S.lycPI203232_1	SAMEA23407	ERS398467	38.78
ERR418072	Solanum lycopersicum	domesticated	S.lycPI311117_1	SAMEA23407	ERS398468	37.67
ERR418073	Solanum lycopersicum	domesticated	S.lycLA1324_1	SAMEA23407	ERS398469	32.97
ERR418074	Solanum lycopersicum	domesticated	S.lycPI158760_1	SAMEA23407	ERS398470	38.03
ERR418075	Solanum lycopersicum	domesticated	S.lycLA0113_1	SAMEA23408	ERS398471	33.94
ERR418076	Solanum lycopersicum	domesticated	S.lycLYC1410_1	SAMEA23408	ERS398472	34.87
ERR418077	Solanum lycopersicum	domesticated	S.lycPI169588_1	SAMEA23408	ERS398473	37.16
ERR418078	Solanum lycopersicum	domesticated	S.lycLYC2962_1	SAMEA23408	ERS398474	37.10
ERR418079	Solanum lycopersicum	domesticated	S.lycLYC2910_1	SAMEA23408	ERS398475	37.61
ERR418080	Solanum pimpinellifolium	wild	S.pimLYC2798_1	SAMEA23408	ERS398476	35.30
ERR418081	Solanum pimpinellifolium	wild	S.pimLYC2740_1	SAMEA23408	ERS398477	37.16
ERR418082	Solanum pimpinellifolium	wild	S.pimLA1584_1	SAMEA23408	ERS398478	32.23
ERR418083	Solanum pimpinellifolium	wild	S.pimLA1578_1	SAMEA23408	ERS398479	36.69
ERR418084	Solanum peruvianum	wild	S.perLA1278_1	SAMEA23408	ERS398480	36.15
ERR418085	Solanum chmielewskii	wild	S.chmLA2663_1	SAMEA23408	ERS398481	35.54
ERR418086	Solanum chmielewskii	wild	S.chmLA2695_1	SAMEA23408	ERS398482	36.50
ERR418087	Solanum cheesmaniae	wild	S.cheLA0483_1	SAMEA23408	ERS398483	34.95
ERR418088	Solanum lycopersicum	domesticated	S.lycCGN15820_1	SAMEA23408	ERS398484	36.59
ERR418089	Solanum cheesmaniae	wild	S.cheLA1401_1	SAMEA23408	ERS398485	37.75
ERR418090	Solanum neorickii	wild	S.neoLA2133_1	SAMEA23408	ERS398486	37.71
ERR418091	Solanum neorickii	wild	S.neoLA0735_1	SAMEA23408	ERS398487	37.11
ERR418092	Solanum arcanum	wild	S.arcLA2157_1	SAMEA23408	ERS398488	33.90
ERR418093	Solanum arcanum	wild	S.arcLA2172_1	SAMEA23408	ERS398489	36.29
ERR418094	Solanum peruvianum	wild	S.perLA1954_1	SAMEA23408	ERS398490	37.51
ERR418095	Solanum huaylasense	wild	S.huaLA1983_1	SAMEA23408	ERS398491	37.67
ERR418096	Solanum huaylasense	wild	S.huaLA1365_1	SAMEA23408	ERS398492	33.44
ERR418097	Solanum chilense	wild	S.chiCGN15532_1	SAMEA23408	ERS398493	33.85
ERR418098	Solanum chilense	wild	S.chiCGN15530_1	SAMEA23408	ERS398494	34.74
ERR418099	Solanum habrochaites	wild	S.habCGN157591_1	SAMEA23408	ERS398495	35.68
ERR418100	Solanum habrochaites	wild	S.habPI134418_1	SAMEA23408	ERS398496	35.97
ERR418101	Solanum habrochaites	wild	S.habCGN157592_1	SAMEA23408	ERS398497	35.46
ERR418102	Solanum habrochaites	wild	S.habLA1718_1	SAMEA23408	ERS398498	35.24
ERR418103	Solanum habrochaites	wild	S.habLA1777_1	SAMEA23408	ERS398499	29.39
ERR418104	Solanum habrochaites	wild	S.habLA0407_1	SAMEA23408	ERS398500	37.76
ERR418105	Solanum habrochaites	wild	S.habLYC4_1	SAMEA23408	ERS398501	33.68
ERR418106	Solanum pennellii	wild	S.penLA1272_1	SAMEA23408	ERS398502	35.66
ERR418107	Solanum pennellii	wild	S.penLA0716_1	SAMEA23408	ERS398503	26.43
ERR418108	Solanum huaylasense	wild	S.huaLA1364_1	SAMEA23408	ERS398504	36.00
ERR418110	Solanum lycopersicum	domesticated	S.lycEA00940_1	SAMEA23408	ERS398506	35.76
ERR418111	Solanum lycopersicum	domesticated	S.lycTR00019_1	SAMEA23408	ERS398507	32.74
ERR418112	Solanum lycopersicum	domesticated	S.lycEA01019_1	SAMEA23408	ERS398508	33.30
ERR418113	Solanum lycopersicum	domesticated	S.lycTR00020_1	SAMEA23408	ERS398509	37.99
ERR418114	Solanum lycopersicum	domesticated	S.lycEA01037_1	SAMEA23408	ERS398510	34.76
ERR418115	Solanum lycopersicum	domesticated	S.lycTR00021_1	SAMEA23408	ERS398511	35.69
ERR418116	Solanum lycopersicum	domesticated	S.lycTR00022_1	SAMEA23408	ERS398512	32.15
ERR418117	Solanum lycopersicum	domesticated	S.lycTR00023_1	SAMEA23408	ERS398513	35.82
ERR418118	Solanum lycopersicum	domesticated	S.lycEA01640_1	SAMEA23408	ERS398514	34.62
ERR418119	Solanum lycopersicum	domesticated	S.lycLA4133_1	SAMEA23408	ERS398515	31.60
ERR418120	Solanum lycopersicum	domesticated	S.lycLA1421_1	SAMEA23408	ERS398516	35.80
ERR418121	Solanum galapagense	wild	S.galLA1044_1	SAMEA23408	ERS398517	32.75
ERR418122	Solanum lycopersicum	domesticated	S.lycLA1479_1	SAMEA23408	ERS398518	35.83

Supplemental Table 1.1. Summary of 82 tomato accession



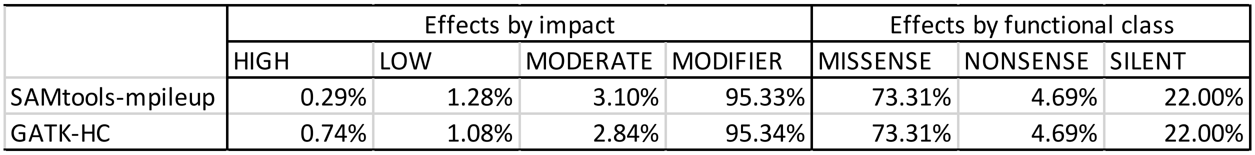
	Domesticated tomatoes				Wild relatives			
	Avg. Percentage	Std. Percentage	Avg. Time	Std. Time	Avg. Percentage	Std. Percentage	Avg. Time	Std. Time
SOAP2	91.24634615	3.776018771	2.488141154	0.444223415	40.577	24.25057044	4.35389	0.539576886
SOAP2-tunec	93.62365385	2.732785714	3.8355775	3.162446692	50.52066667	21.69383687	11.32610833	3.162446692
Bowtie2	97.9325	1.067468831	16.06570769	1.789843112	79.16666667	10.08325732	17.49555	1.213827365
Bowtie2-tun	98.26	0.928536694	23.55064423	1.840175376	83.32566667	7.972755195	27.93500333	2.60995463
BWA-MEM	99.54384615	0.342597654	10.27083308	1.561096856	95.946	1.907017169	23.75333333	6.542826671

Supplemental Table 1.2. Summary of alignment time



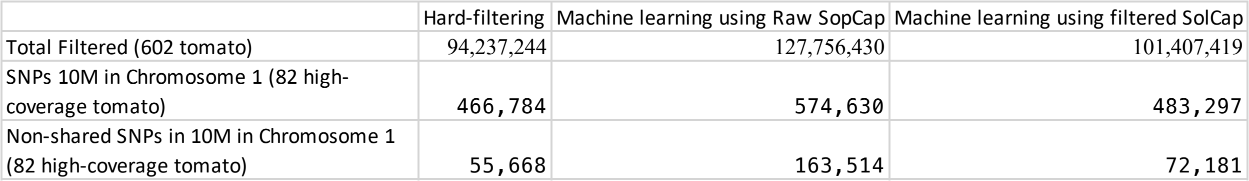
	Total	Syntentic - Under Size	Syntenic - Over Size	10kb- Under Size	10kb - Over Size	>10kb - Under Size	> 10kb - Ove	DiffChr - Unc	DiffChr - Ove
ERR418070_bowtie2	27074	12660	10818	0	0	55	6	1757	1778
ERR418070_bwa	104996	39944	29646	1	0	4331	907	17495	12672
ERR418095_bowtie2	19848	9579	8152	0	0	26	2	1082	1007
ERR418095_bwa	77146	30796	21998	0	0	2769	888	11985	8710
ERR418105_bowtie2	14435	7173	5849	0	0	48	0	724	641
ERR418105_bwa	73111	28898	19284	2	0	3191	1331	12274	8131
ERR418106_bowtie2	17251	8350	7034	0	0	36	1	994	836
ERR418106_bwa	73781	28144	21015	0	0	2920	1149	11795	8758
ERR418107_bowtie2	17686	8462	6920	0	0	80	1	1162	1061
ERR418107_bwa	85534	30804	20964	1	0	5174	2391	15346	10854
ERR418118_bowtie2	26988	12532	10835	0	0	60	5	1833	1723
ERR418118_bwa	106180	40855	29759	0	0	4339	766	17735	12726

Supplemental Table 1.3. Summary of synteny analysis



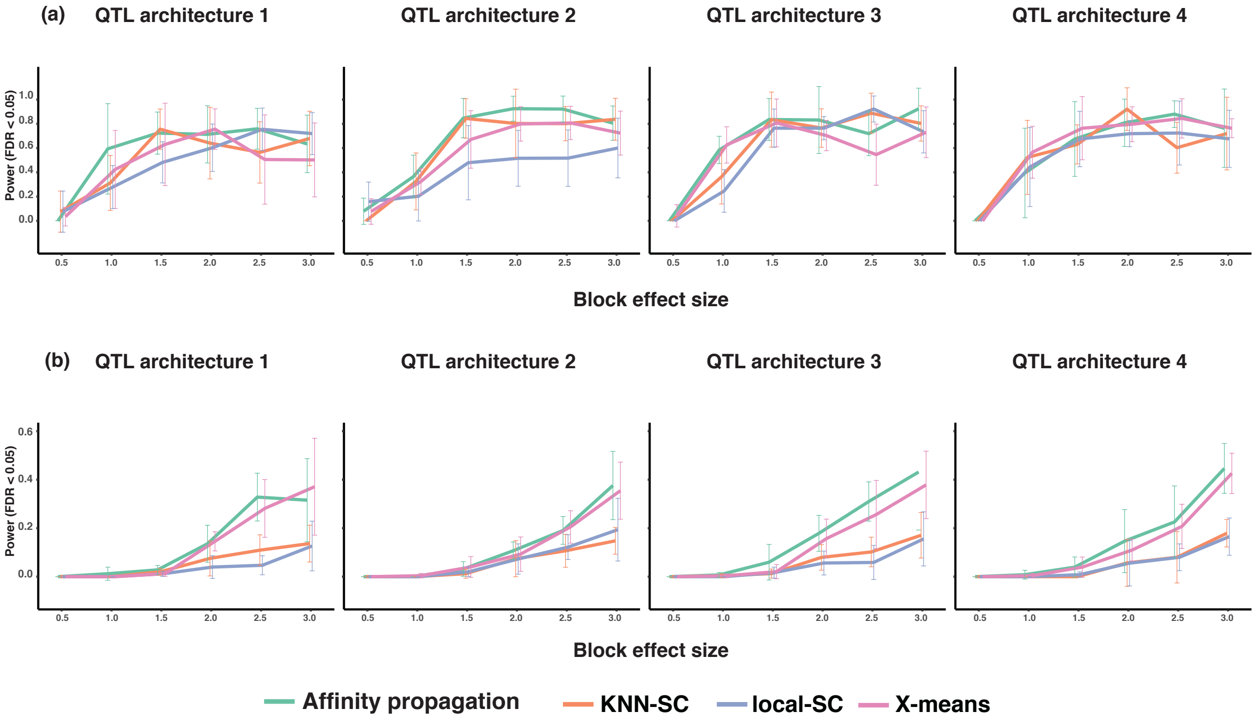
	Effects by impact				Effects by functional class		
	HIGH	LOW	MODERATE	MODIFIER	MISSENSE	NONSENSE	SILENT
SAMtools-mpileup	0.29%	1.28%	3.10%	95.33%	73.31%	4.69%	22.00%
GATK-HC	0.74%	1.08%	2.84%	95.34%	73.31%	4.69%	22.00%

Supplemental Table 1.4. Functional annotation summary of variants identified by different variant calling programs



	Hard-filtering	Machine learning using Raw SopCap	Machine learning using filtered SolCap
Total Filtered (602 tomato)	94,237,244	127,756,430	101,407,419
SNPs 10M in Chromosome 1 (82 high-coverage tomato)	466,784	574,630	483,297
Non-shared SNPs in 10M in Chromosome 1 (82 high-coverage tomato)	55,668	163,514	72,181

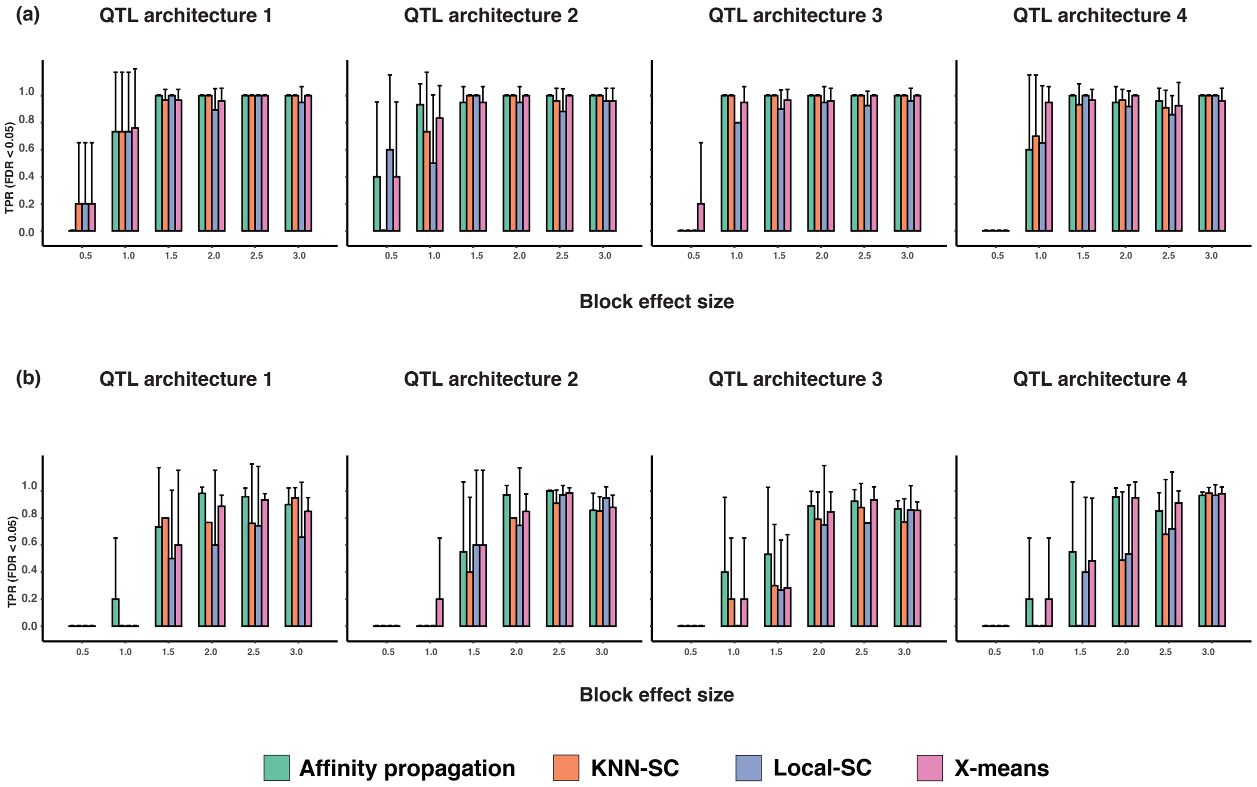
Supplemental Table 1.5. Summary of different variant filtering results



Supplemental Figure 2.1. Mapping power comparison of different haplotype clustering algorithms.

(a) Mapping power (FDR < 0.05) of affinity propagation, KNN-spectral clustering, local-spectral clustering and X-means in the low polygenicity simulation

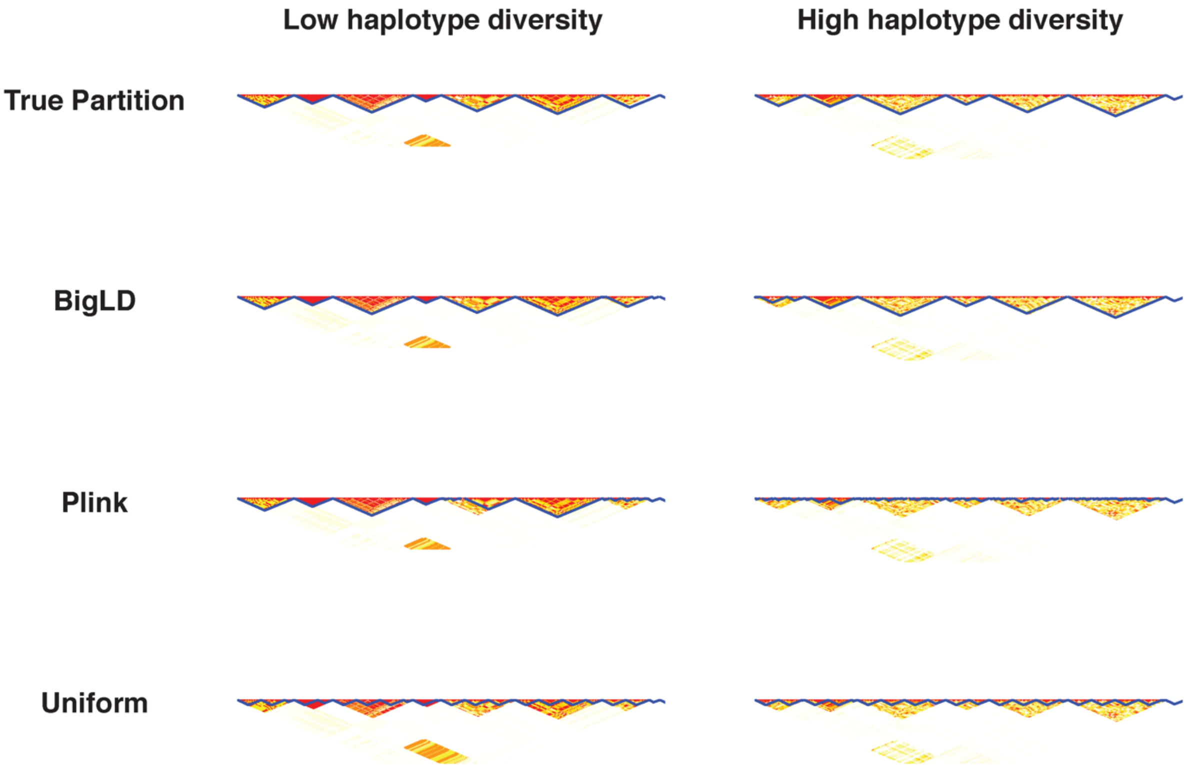
(b) Mapping power (FDR < 0.05) of affinity propagation, KNN-spectral clustering, local-spectral clustering and X-means in the high polygenicity simulation



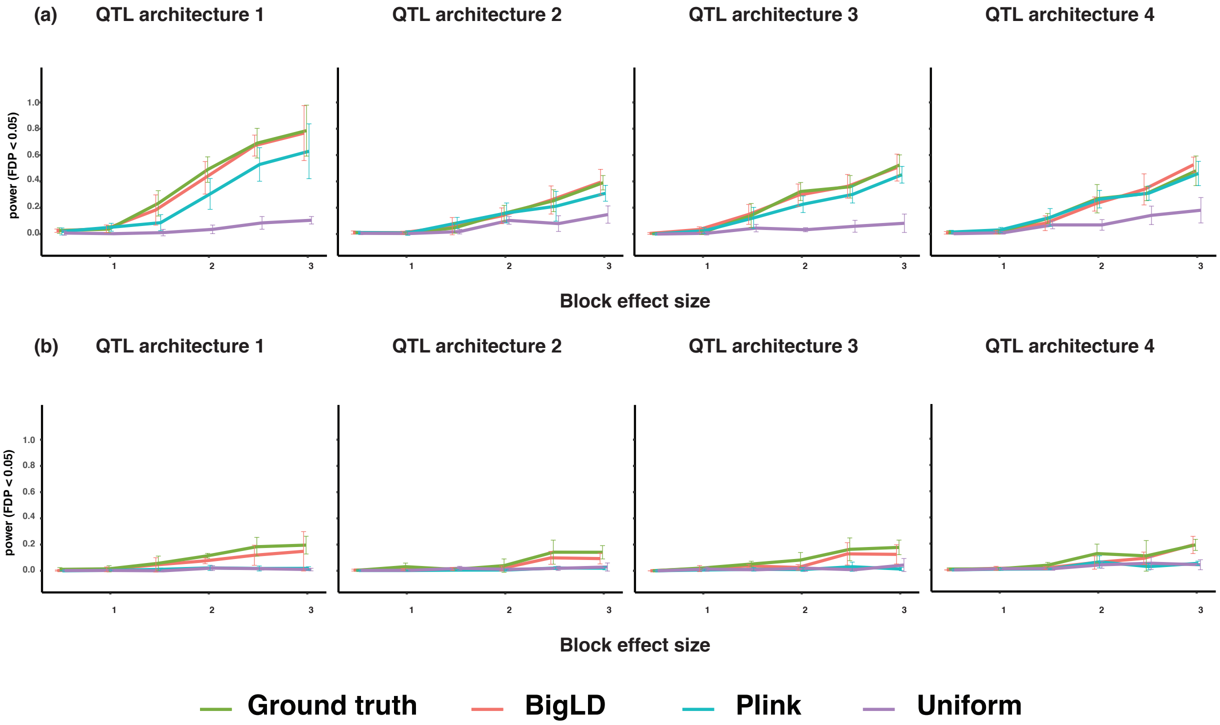
Supplemental Figure 2.2. True positive rate (TPR) comparison of different haplotype clustering algorithms.

(a) TPR (FDR < 0.05) of affinity propagation, KNN-spectral clustering, local-spectral clustering and X-means in the low polygenicity simulation

(b) TPR (FDR < 0.05) of affinity propagation, KNN-spectral clustering, local-spectral clustering and X-means in the high polygenicity simulation



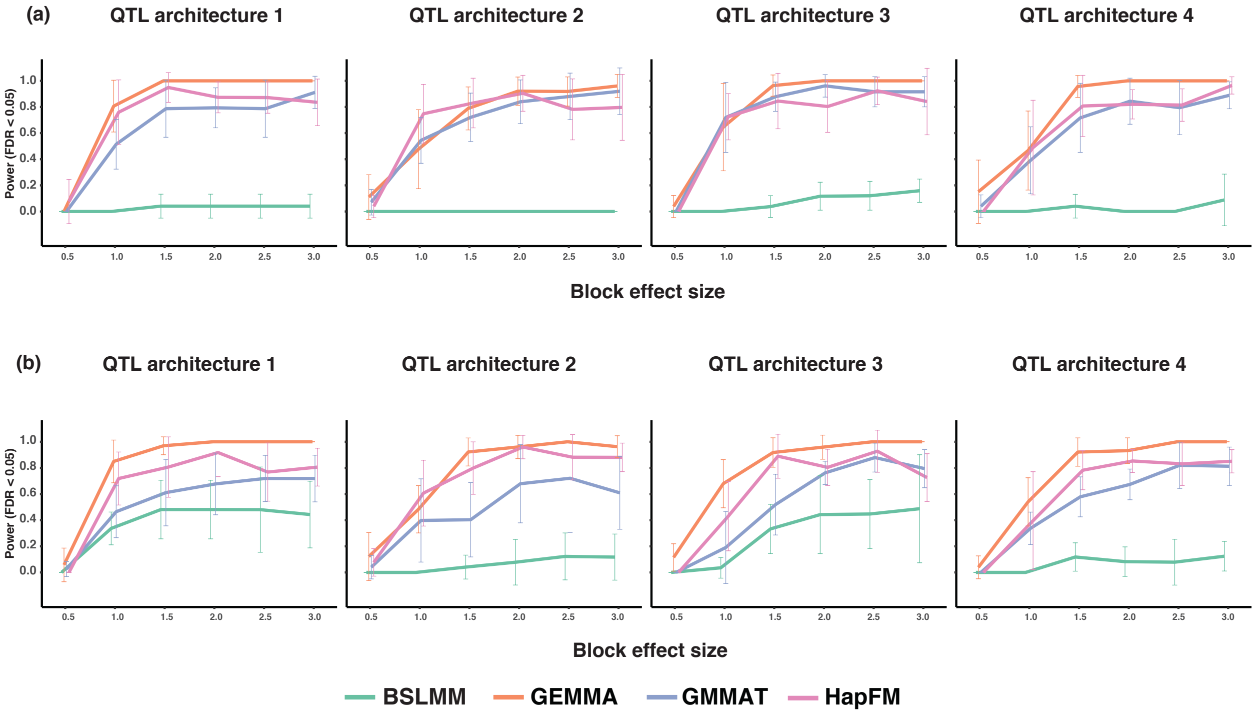
Supplemental Figure 2.3. Comparison of Block partition algorithms on simulated datasets. Each block partition algorithm was tested on low haplotype diversity and high haplotype diversity simulations. The redness indicates the strength of LD between SNP pairs, and the blue line indicates the block partition generated by the method.



Supplemental Figure 2.4. Mapping power comparison of different block partition algorithms in the low polygenicity simulations. The x-axis indicates the per-locus heritability.

(b). Mapping power comparison (FDR < 0.05) of block partition algorithms in the low haplotype diversity and low polygenicity simulations.

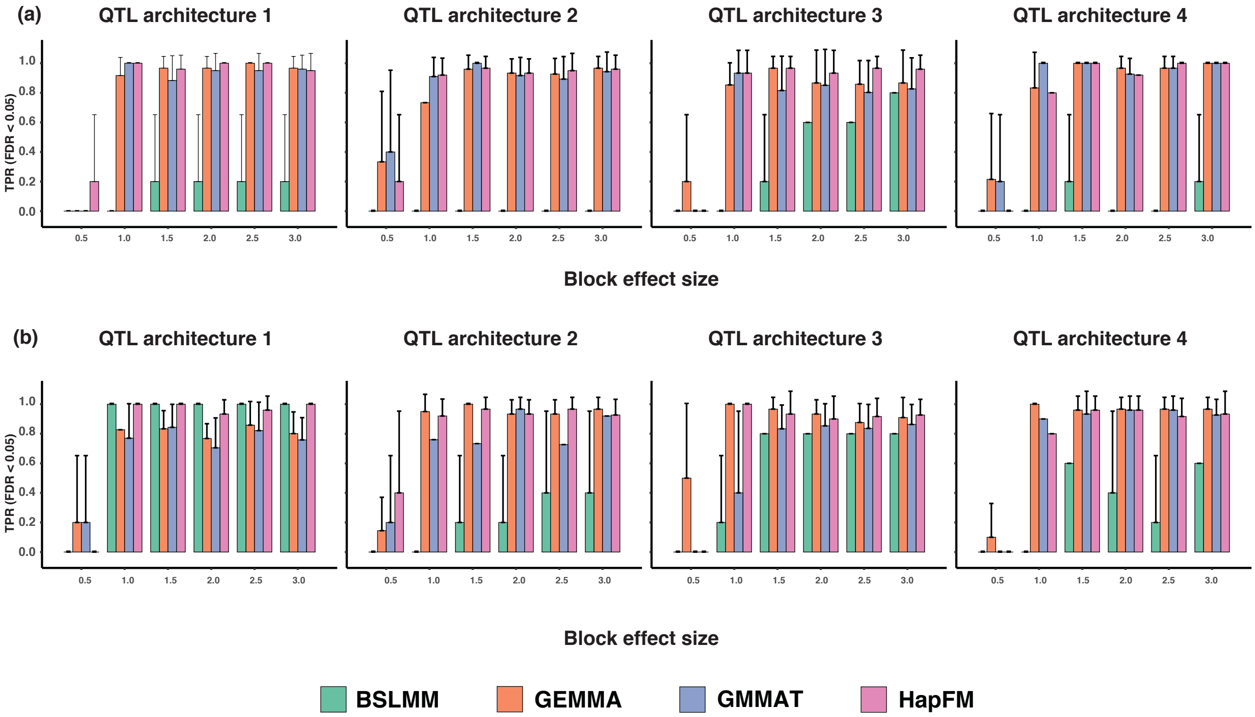
(c). Mapping power comparison (FDR < 0.05) of block partition algorithms in the high haplotype diversity and low polygenicity simulations.



Supplemental Figure 2.5. Mapping power comparison of different GWAS algorithms in the low polygenicity simulations. The x-axis indicates the per-locus heritability.

(a). Mapping power comparison (FDR < 0.05) of different GWAS algorithms in the low haplotype diversity and low polygenicity simulations.

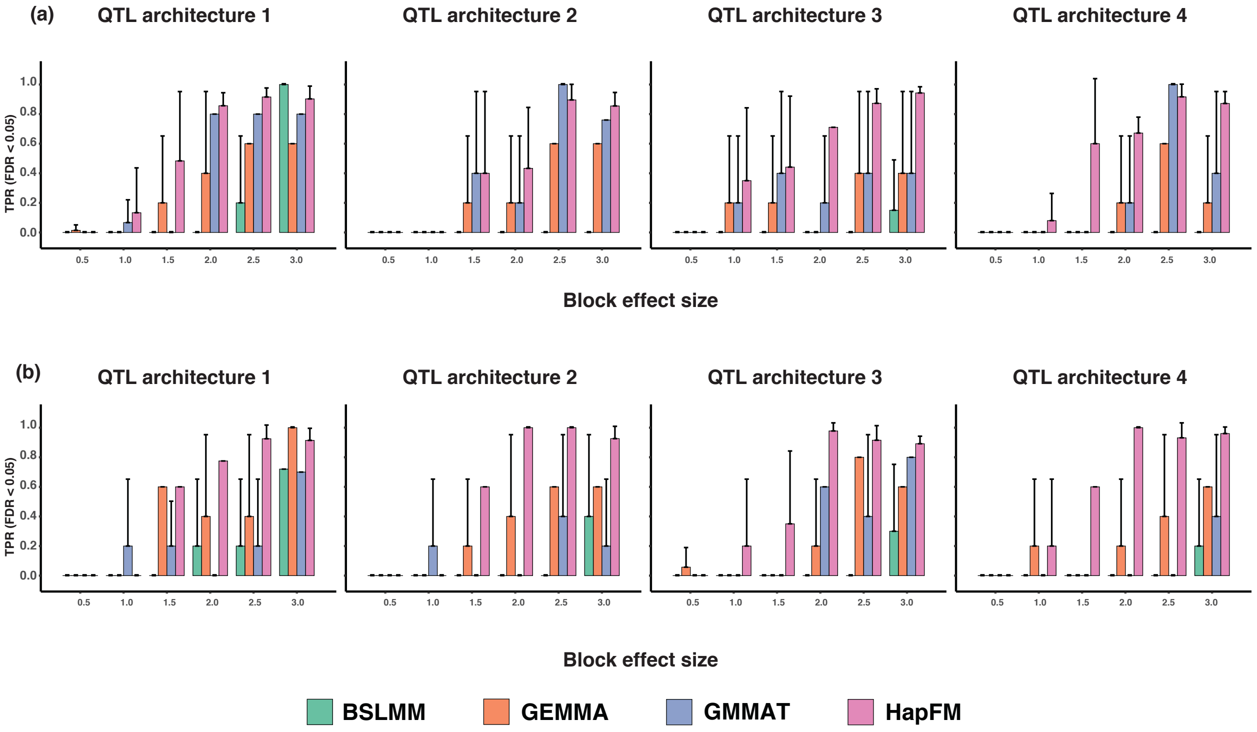
(b). Mapping power comparison (FDR < 0.05) of different GWAS algorithms in the high haplotype diversity and low polygenicity simulations.



Supplemental Figure 2.6. True positive rate different GWAS algorithms in the low polygenicity simulations. The x-axis indicates the per-locus heritability.

(a). True positive rate (FDR < 0.05) of different GWAS algorithms in the low haplotype diversity and low polygenicity simulations.

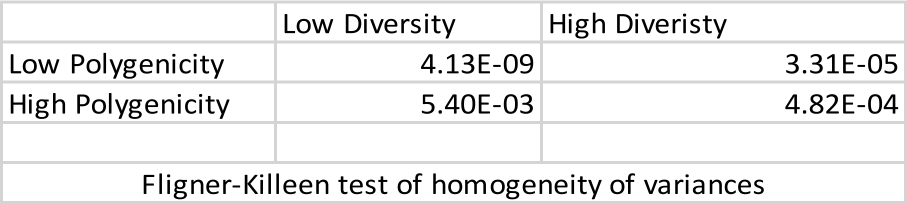
(b). True positive rate (FDR < 0.05) of different GWAS algorithms in the high haplotype diversity and low polygenicity simulations.



Supplemental Figure 2.7. True positive rate different GWAS algorithms in the high polygenicity simulations. The x-axis indicates the per-locus heritability.

(a). True positive rate (FDR < 0.05) of different GWAS algorithms in the low haplotype diversity and low polygenicity simulations.

(b). True positive rate (FDR < 0.05) of different GWAS algorithms in the high haplotype diversity and low polygenicity simulations.



	Low Diversity	High Diveristy
Low Polygenicity	4.13E-09	3.31E-05
High Polygenicity	5.40E-03	4.82E-04
Fligner-Killeen test of homogeneity of variances		

Supplemental Table 2.1. Fligner-Killeen test of the variance of mapping interval length between GEMMA and other GWAS methods.

	HapFM	FH-GWAS	GEMMA	GMMAT	BSLMM
Previously identified loci (genes)					
FT	Y	Y	Y	N	Y
FLC	N	Y	Y	N	N
DOG1	Y	Y	Y	Y	N
VIN3	Y	Y	Y	Y	N
Novel loci					
Chr3@7598564-7598957	Y	N	N	N	N
Chr4@405136-406621	Y	Y	N	N	N
Chr5@14063228-14197451	Y	N	N	N	N
Chr5@16141604-16146257	Y	N	N	N	N

Supplemental Table 2.2. Comparison of GWAS methods using Arabidopsis FT10 dataset

Bibliography

1. Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., *et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

2. Albert, J.H., and Chib, S. (1993). Bayesian-Analysis of Binary and Polychotomous Response Data. J Am Stat Assoc *88*, 669-679.

3. Albertin, C.B., Simakov, O., Mitros, T., Wang, Z.Y., Pungor, J.R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C.W., and Rokhsar, D.S. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature *524*, 220-224.

4. Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., *et al.* (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell *182*, 145-161 e123.

5. Anderson, J.E., Kantar, M.B., Kono, T.Y., Fu, F., Stec, A.O., Song, Q., Cregan, P.B., Specht, J.E., Diers, B.W., Cannon, S.B., *et al.* (2014). A roadmap for functional structural variants in the soybean genome. G3 (Bethesda) *4*, 1307-1318.

6. Arabidopsis Genome, I. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796-815.

7. Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-Briere, C., Owens, G.L., Carrere, S., Mayjonade, B., *et al.* (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature *546*, 148-152.

8. Bayer, P.E., Golicz, A.A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. Nat Plants *6*, 914-920.

9. Bevan, M.W., Uauy, C., Wulff, B.B., Zhou, J., Krasileva, K., and Clark, M.D. (2017). Genomic innovation for crop improvement. Nature *543*, 346-354.

10. Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J Stat Mech-Theory E.

11. Bolger, A., Scossa, F., Bolger, M.E., Lanz, C., Maumus, F., Tohge, T., Quesneville, H., Alseekh, S., Sorensen, I., Lichtenstein, G., *et al.* (2014a). The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat Genet *46*, 1034-1038.

12. Bolger, A.M., Lohse, M., and Usadel, B. (2014b). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

13. Borile, C., Labarre, M., Franz, S., Sola, C., and Refregier, G. (2011). Using affinity propagation for identifying subspecies among clonal organisms: lessons from M. tuberculosis. BMC Bioinformatics *12*, 224.

14. Brachi, B., Morris, G.P., and Borevitz, J.O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol *12*, 232.

15. Browning, B.L., and Browning, S.R. (2016). Genotype Imputation with Millions of Reference Samples. Am J Hum Genet *98*, 116-126.

16. Browning, S.R., and Browning, B.L. (2007). 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*, 1084-1097.

17. Brzyski, D., Peterson, C.B., Sobczyk, P., Candes, E.J., Bogdan, M., and Sabatti, C. (2017). Controlling the Rate of GWAS False Discoveries. Genetics *205*, 61-75.

18. Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C.*, et al*. (2009). The genetic architecture of maize flowering time. Science *325*, 714-718.

19. Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B., Xu, D., Yang, B., Xie, C.*, et al*. (2018). Construction of the third-generation Zea mays haplotype map. Gigascience *7*, 1-12.

20. Callaway, E. (2014). Domestication: The birth of rice. Nature *514*, S58-59.

21. Chan, Y., Salem, R.M., Hsu, Y.H., McMahon, G., Pers, T.H., Vedantam, S., Esko, T., Guo, M.H., Lim, E.T., Consortium, G.*, et al*. (2015). 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.

22. Chen, H., Huffman, J.E., Brody, J.A., Wang, C., Lee, S., Li, Z., Gogarten, S.M., Sofer, T., Bielak, L.F., Bis, J.C.*, et al*. (2019a). 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.

23. Chen, J., Li, X., Zhong, H., Meng, Y., and Du, H. (2019b). Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Sci Rep *9*, 9345.

24. Chen, K., Wang, Y., Zhang, R., Zhang, H., and Gao, C. (2019c). CRISPR/Cas Genome Editing and Precision Plant Breeding in Agriculture. Annu Rev Plant Biol *70*, 667-697.

25. Cheng, A.Y., Teo, Y.Y., and Ong, R.T. (2014). Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics *30*, 1707-1713.

26. Choi, J.Y., Lye, Z.N., Groen, S.C., Dai, X., Rughani, P., Zaaijer, S., Harrington, E.D., Juul, S., and Purugganan, M.D. (2020). Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. Genome Biol *21*, 21.

27. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80-92.

28. Clevenger, J., Chavarro, C., Pearl, S.A., Ozias-Akins, P., and Jackson, S.A. (2015). Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. Mol Plant *8*, 831-846.

29. Computational Pan-Genomics, C. (2018). Computational pan-genomics: status, promises and challenges. Brief Bioinform *19*, 118-135.

30. Cook, D.E., Lee, T.G., Guo, X., Melito, S., Wang, K., Bayless, A.M., Wang, J., Hughes, T.J., Willis, D.K., Clemente, T.E.*, et al.* (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science *338*, 1206-1209.

31. Cortes, L.T., Zhang, Z.W., and Yu, J.M. (2021). Status and prospects of genome-wide association studies in plants. Plant Genome *14*.

32. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T.*, et al.* (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156-2158.

33. Danecek, P., and McCarthy, S.A. (2017). BCFtools/csq: haplotype-aware variant consequences. Bioinformatics *33*, 2037-2039.

34. Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M.*, et al.* (2016). Next-generation genotype imputation service and methods. Nat Genet *48*, 1284-1287.

35. De Summa, S., Malerba, G., Pinto, R., Mori, A., Mijatovic, V., and Tommasi, S. (2017). GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC Bioinformatics *18*, 119.

36. Della Coletta, R., Qiu, Y., Ou, S., Hufford, M.B., and Hirsch, C.N. (2021). How the pan-genome is changing crop genomics and improvement. Genome Biol *22*, 3.

37. Ding, K., and Kullo, I.J. (2007). Methods for the selection of tagging SNPs: a comparison of tagging efficiency and performance. Eur J Hum Genet *15*, 228-236.

38. Doebley, J.F., Gaut, B.S., and Smith, B.D. (2006). The molecular genetics of crop domestication. Cell *127*, 1309-1321.

39. Du, H., Yu, Y., Ma, Y., Gao, Q., Cao, Y., Chen, Z., Ma, B., Qi, M., Li, Y., Zhao, X.*, et al.* (2017). Sequencing and de novo assembly of a near complete indica rice genome. Nat Commun *8*, 15324.

40. Farashi, S., Kryza, T., Clements, J., and Batra, J. (2019). Post-GWAS in prostate cancer: from genetic association to biological contribution. Nat Rev Cancer *19*, 46-59.

41. Ferreiro-Iglesias, A., Lesseur, C., McKay, J., Hung, R.J., Han, Y., Zong, X., Christiani, D., Johansson, M., Xiao, X., Li, Y.*, et al.* (2018). Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. Nat Commun *9*, 3927.

42. Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. Science *315*, 972-976.

43. Fuentes, R.R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J.F., Mohiyuddin, M., Wing, R.A., McNally, K.L., Tatarinova, T., Grigoriev, A.*, et al.* (2019). Structural variants in 3000 rice genomes. Genome Res *29*, 870-880.

44. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., *et al.* (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225-2229.

45. Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The Post-GWAS Era: From Association to Function. Am J Hum Genet *102*, 717-730.

46. Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D.M., Burzynski-Chang, E.A., Fish, T.L., Stromberg, K.A., Sacks, G.L., *et al.* (2019a). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet.

47. Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D.M., Burzynski-Chang, E.A., Fish, T.L., Stromberg, K.A., Sacks, G.L., *et al.* (2019b). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet *51*, 1044-1051.

48. Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv *1207.3907*.

49. Genomes Consortium. Electronic address, m.n.g.o.a.a., and Genomes, C. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell *166*, 481-491.

50. Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

51. Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., *et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science *296*, 92-100.

52. Goktay, M., Fulgione, A., and Hancock, A.M. (2021). 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.

53. Golicz, A.A., Batley, J., and Edwards, D. (2016). Towards plant pangenomics. Plant Biotechnol J *14*, 1099-1105.

54. Gross, B.L., and Olsen, K.M. (2010). Genetic perspectives on crop domestication. Trends in Plant Science *15*, 529-537.

55. Haberer, G., Kamal, N., Bauer, E., Gundlach, H., Fischer, I., Seidel, M.A., Spannagl, M., Marcon, C., Ruban, A., Urbany, C., *et al.* (2020). European maize genomes highlight intraspecies variation in repeat and gene content. Nat Genet *52*, 950-957.

56. Hamazaki, K., and Iwata, H. (2020). RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. PLoS Comput Biol *16*, e1007663.

57. Heslop-Harrison, J.S.P., and Schwarzacher, T. (2012). Genetics and genomics of crop domestication. Plant Biotechnology and Agriculture: Prospects for the 21st Century, 3-18.

58. Highnam, G., Wang, J.J., Kusler, D., Zook, J., Vijayan, V., Leibovich, N., and Mittelman, D. (2015). An analytical framework for optimizing variant discovery from personal genomes. Nat Commun *6*, 6275.

59. Holland, J.B. (2007). Genetic architecture of complex traits in plants. Curr Opin Plant Biol *10*, 156-161.

60. Horton, M.W., Hancock, A.M., Huang, Y.S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N.W., Platt, A., Sperone, F.G., Vilhjalmsson, B.J*., et al*. (2012). Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nat Genet *44*, 212-216.

61. Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L*., et al*. (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature *496*, 498-503.

62. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet *5*, e1000529.

63. Huang, X., and Han, B. (2014a). Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol *65*, 531-551.

64. Huang, X.H., and Han, B. (2014b). Natural Variations and Genome-Wide Association Studies in Crop Plants. Annual Review of Plant Biology, Vol 65 *65*, 531-551.

65. Huang, Y., Huang, W., Meng, Z., Braz, G.T., Li, Y., Wang, K., Wang, H., Lai, J., Jiang, J., Dong, Z*., et al*. (2021). Megabase-scale presence-absence variation with Tripsacum origin was under selection during maize domestication and adaptation. Genome Biol *22*, 237.

66. Hubner, S., Bercovich, N., Todesco, M., Mandel, J.R., Odenheimer, J., Ziegler, E., Lee, J.S., Baute, G.J., Owens, G.L., Grassa, C.J*., et al*. (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat Plants *5*, 54-62.

67. Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J., Ricci, W.A., Guo, T., Olson, A., Qiu, Y*., et al*. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science *373*, 655-662.

68. Hufford, M.B., Xu, X., van Heerwaarden, J., Pyhajarvi, T., Chia, J.M., Cartwright, R.A., Elshire, R.J., Glaubitz, J.C., Guill, K.E., Kaeppler, S.M*., et al*. (2012). Comparative population genomics of maize domestication and improvement. Nat Genet *44*, 808-811.

69. Hwang, E.Y., Song, Q., Jia, G., Specht, J.E., Hyten, D.L., Costa, J., and Cregan, P.B. (2014). A genome-wide association study of seed protein and oil content in soybean. BMC Genomics *15*, 1.

70. Ingvarsson, P.K., and Street, N.R. (2011). Association genetics of complex traits in plants. New Phytol *189*, 909-922.

71. International HapMap, C., Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P*., et al*. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

72. Jacob, P., Avni, A., and Bendahmane, A. (2018). Translational Research: Exploring and Creating Genetic Diversity. Trends Plant Sci *23*, 42-52.

73. Jaganathan, D., Bohra, A., Thudi, M., and Varshney, R.K. (2020). Fine mapping and gene cloning in the post-NGS era: advances and prospects. Theoretical and Applied Genetics *133*, 1791-1810.

74. Jakoby, M.J., Falkenhan, D., Mader, M.T., Brininstool, G., Wischnitzki, E., Platz, N., Hudson, A., Hulskamp, M., Larkin, J., and Schnittger, A. (2008). Transcriptional profiling of mature Arabidopsis trichomes reveals that NOECK encodes the MIXTA-like transcriptional regulator MYB106. Plant Physiol *148*, 1583-1602.

75. Kawakatsu, T., Huang, S.C., Jupe, F., Sasaki, E., Schmitz, R.J., Urich, M.A., Castanon, R., Nery, J.R., Barragan, C., He, Y*., et al.* (2016). Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell *166*, 492-505.

76. Kim, J.E., Oh, S.K., Lee, J.H., Lee, B.M., and Jo, S.H. (2014). Genome-wide SNP calling using next generation sequencing data in tomato. Mol Cells *37*, 36-42.

77. Kim, M.S., Lee, T., Baek, J., Kim, J.H., Kim, C., and Jeong, S.C. (2021). Genome assembly of the popular Korean soybean cultivar Hwangkeum. G3 (Bethesda) *11*.

78. Kim, S.A., Cho, C.S., Kim, S.R., Bull, S.B., and Yoo, Y.J. (2018). 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.

79. Klepikova, A.V., Kasianov, A.S., Gerasimov, E.S., Logacheva, M.D., and Penin, A.A. (2016). A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant J *88*, 1058-1070.

80. Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. Plant Methods *9*, 29.

81. Krause, K., Johnsen, H.R., Pielach, A., Lund, L., Fischer, K., and Rose, J.K.C. (2018). Identification of tomato introgression lines with enhanced susceptibility or resistance to infection by parasitic giant dodder (Cuscuta reflexa). Physiol Plant *162*, 205-218.

82. Krishnan, V., Utiramerur, S., Ng, Z., Datta, S., Snyder, M.P., and Ashley, E.A. (2021). Benchmarking workflows to assess performance and suitability of germline variant calling pipelines in clinical diagnostic assays. Bmc Bioinformatics *22*.

83. Kumaran, M., Subramanian, U., and Devarajan, B. (2019). Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. Bmc Bioinformatics *20*.

84. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol *5*, R12.

85. Lam, H.M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.L., Li, M.W., He, W., Qin, N., Wang, B*., et al.* (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet *42*, 1053-1059.

86. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

87. Lei, L., Goltsman, E., Goodstein, D., Wu, G.A., Rokhsar, D.S., and Vogel, J.P. (2021). Plant Pan-Genomics Comes of Age. Annu Rev Plant Biol *72*, 411-435.

88. Li, B., Gao, J., Chen, J., Wang, Z., Shen, W., Yi, B., Wen, J., Ma, C., Shen, J., Fu, T.*, et al.* (2020). Identification and fine mapping of a major locus controlling branching in Brassica napus. Theor Appl Genet *133*, 771-783.

89. Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R.J., Ma, Z., Shang, H., Ma, X., Wu, J.*, et al.* (2015). Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. Nat Biotechnol *33*, 524-530.

90. Li, H. (2015). FermiKit: assembly-based variant calling for Illumina resequencing data. Bioinformatics *31*, 3694-3696.

91. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589-595.

92. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

93. Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform *11*, 473-483.

94. Li, N., Lin, B., Wang, H., Li, X., Yang, F., Ding, X., Yan, J., and Chu, Z. (2019). Natural variation in ZmFBL41 confers banded leaf and sheath blight resistance in maize. Nat Genet *51*, 1540-1548.

95. Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009b). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966-1967.

96. Liang, Y., He, L., Zhao, Y.R., Hao, Y.Y., Zhou, Y.F., Li, M.L., Li, C., Pu, X.M., and Wen, Z.N. (2019). Comparative Analysis for the Performance of Variant Calling Pipelines on Detecting the de novo Mutations in Humans. Front Pharmacol *10*.

97. Lin, G., He, C., Zheng, J., Koo, D.H., Le, H., Zheng, H., Tamang, T.M., Lin, J., Liu, Y., Zhao, M.*, et al.* (2021). Chromosome-level genome assembly of a regenerable maize inbred line A188. Genome Biol *22*, 175.

98. Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X.*, et al.* (2014). Genomic analyses provide insights into the history of tomato breeding. Nat Genet *46*, 1220-1226.

99. Liu, F., Schmidt, R.H., Reif, J.C., and Jiang, Y. (2019). Selecting Closely-Linked SNPs Based on Local Epistatic Effects for Haplotype Construction Improves Power of Association Mapping. G3 (Bethesda) *9*, 4115-4126.

100. Liu, X., Han, S., Wang, Z., Gelernter, J., and Yang, B.Z. (2013). Variant callers for next-generation sequencing data: a comparison study. PLoS One *8*, e75619.

101. Liu, X., Huang, M., Fan, B., Buckler, E.S., and Zhang, Z. (2016). Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLoS Genet *12*, e1005767.

102. Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z., Shi, M.*, et al.* (2020). Pan-Genome of Wild and Cultivated Soybeans. Cell *182*, 162-176 e113.

103. Lv, Q., Li, W., Sun, Z., Ouyang, N., Jing, X., He, Q., Wu, J., Zheng, J., Zheng, J., Tang, S.*, et al.* (2020a). Resequencing of 1,143 indica rice accessions reveals important genetic variations and different heterosis patterns. Nat Commun *11*, 4778.

104. Lv, Q.M., Li, W.G., Sun, Z.Z., Ouyang, N., Jing, X., He, Q., Wu, J., Zheng, J.K., Zheng, J.T., Tang, S.Q.*, et al.* (2020b). Resequencing of 1,143 indica rice accessions reveals important genetic variations and different heterosis patterns. Nature Communications *11*.

105. Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat Rev Genet *11*, 31-46.

106. Michael, T.P., and VanBuren, R. (2015). Progress, challenges and the future of crop genomes. Curr Opin Plant Biol *24*, 71-81.

107. Migicovsky, Z., and Myles, S. (2017). Exploiting Wild Relatives for Genomics-assisted Breeding of Perennial Crops. Front Plant Sci *8*, 460.

108. Mitchell, M. (1996). An introduction to genetic algorithms (Cambridge, Mass.: MIT Press).

109. Morrell, P.L., Buckler, E.S., and Ross-Ibarra, J. (2011). Crop genomics: advances and applications. Nat Rev Genet *13*, 85-96.

110. Nagarajan, N., and Pop, M. (2013). Sequence assembly demystified. Nat Rev Genet *14*, 157-167.

111. Narsai, R., Law, S.R., Carrie, C., Xu, L., and Whelan, J. (2011). In-depth temporal transcriptome profiling reveals a crucial developmental switch with roles for RNA processing and organelle metabolism that are essential for germination in Arabidopsis. Plant Physiol *157*, 1342-1362.

112. Novikov, A.V. (2019). PyClustering: Data Mining Library. Journal of Open Source Software *4*, 1230.

113. Ogbonna, A.C., Braatz de Andrade, L.R., Rabbi, I.Y., Mueller, L.A., Jorge de Oliveira, E., and Bauchet, G.J. (2021). 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.

114. Pagnussat, G.C., Yu, H.J., Ngo, Q.A., Rajani, S., Mayalagu, S., Johnson, C.S., Capron, A., Xie, L.F., Ye, D., and Sundaresan, V. (2005). Genetic and molecular identification of genes required for female gametophyte development and function in Arabidopsis. Development *132*, 603-614.

115. Paten, B., Novak, A.M., Eizenga, J.M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. Genome Res *27*, 665-676.

116. Peiffer, J.A., Romay, M.C., Gore, M.A., Flint-Garcia, S.A., Zhang, Z., Millard, M.J., Gardner, C.A., McMullen, M.D., Holland, J.B., Bradbury, P.J., *et al*. (2014). The genetic architecture of maize height. Genetics *196*, 1337-1356.

117. Pelleg, D., and Moore, A. (2000). 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.

118. Periyannan, S., Moore, J., Ayliffe, M., Bansal, U., Wang, X., Huang, L., Deal, K., Luo, M., Kong, X., Bariana, H., *et al*. (2013). The gene Sr33, an ortholog of barley Mla genes, encodes resistance to wheat stem rust race Ug99. Science *341*, 786-788.

119. Pop, M., and Salzberg, S.L. (2008). Bioinformatics challenges of new sequencing technology. Trends Genet *24*, 142-149.

120. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., *et al*. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv.

121. project, r.g. (2014). The 3,000 rice genomes project. Gigascience *3*, 7.

122. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., *et al*. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet *81*, 559-575.

123. Rambla, J.L., Medina, A., Fernandez-Del-Carmen, A., Barrantes, W., Grandillo, S., Cammareri, M., Lopez-Casado, G., Rodrigo, G., Alonso, A., Garcia-Martinez, S., *et al*. (2017). Identification, introgression, and validation of fruit volatile QTLs from a red-fruited wild tomato species. J Exp Bot *68*, 429-442.

124. Reinert, K., Dadi, T.H., Ehrhardt, M., Hauswedell, H., Mehringer, S., Rahn, R., Kim, J., Pockrandt, C., Winkler, J., Siragusa, E., *et al*. (2017). The SeqAn C++ template library for efficient sequence analysis: A resource for programmers. J Biotechnol *261*, 157-168.

125. Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet *19*, 491-504.

126. Schatz, M.C., Witkowski, J., and McCombie, W.R. (2012). Current challenges in de novo plant genome sequencing and assembly. Genome Biol *13*, 243.

127. Schilbert, H.M., Rempel, A., and Pucker, B. (2020). Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. Plants (Basel) *9*.

128. Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., *et al*. (2010). Genome sequence of the palaeopolyploid soybean. Nature *463*, 178-183.

129. Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., *et al*. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112-1115.

130. Segura, V., Vilhjalmsson, B.J., Platt, A., Korte, A., Seren, U., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet *44*, 825-830.

131. Seren, U., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K., and Korte, A. (2017). AraPheno: a public database for Arabidopsis thaliana phenotypes. Nucleic Acids Res *45*, D1054-D1059.

132. Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W., and Shen, B. (2014). Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed Res Int *2014*, 309650.

133. Shinohara, N., Ohbayashi, I., and Sugiyama, M. (2014). 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.

134. Sim, S.C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M.W., Van Deynze, A., Hamilton, J.P., Buell, C.R., Causse, M., Wijeratne, S*., et al*. (2012). Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS One *7*, e40563.

135. Song, Q., Hyten, D.L., Jia, G., Quigley, C.V., Fickus, E.W., Nelson, R.L., and Cregan, P.B. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One *8*, e54985.

136. Strickler, S.R., Bombarely, A., Munkvold, J.D., York, T., Menda, N., Martin, G.B., and Mueller, L.A. (2015). Comparative genomics and phylogenetic discordance of cultivated tomato and close wild relatives. PeerJ *3*, e793.

137. Thankaswamy-Kosalai, S., Sen, P., and Nookaew, I. (2017). Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics *109*, 186-191.

138. Thomson, M.J., Singh, N., Dwiyanti, M.S., Wang, D.R., Wright, M.H., Perez, F.A., DeClerck, G., Chin, J.H., Malitic-Layaoen, G.A., Juanillas, V.M*., et al*. (2017). Large-scale deployment of a rice 6 K SNP array for genetics and breeding applications. Rice (N Y) *10*, 40.

139. Tieman, D., Zhu, G., Resende, M.F., Jr., Lin, T., Nguyen, C., Bies, D., Rambla, J.L., Beltran, K.S., Taylor, M., Zhang, B*., et al*. (2017). A chemical genetic roadmap to improved tomato flavor. Science *355*, 391-394.

140. Tomato Genome, C. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature *485*, 635-641.

141. Tomato Genome Sequencing, C., Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N*., et al*. (2014). Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. Plant J *80*, 136-148.

142. Tran, D.T., Steketee, C.J., Boehm, J.D., Jr., Noe, J., and Li, Z. (2019). Genome-Wide Association Analysis Pinpoints Additional Major Genomic Regions Conferring Resistance to Soybean Cyst Nematode (Heterodera glycines Ichinohe). Front Plant Sci *10*, 401.

143. Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet *13*, 36-46.

144. Valliyodan, B., Cannon, S.B., Bayer, P.E., Shu, S., Brown, A.V., Ren, L., Jenkins, J., Chung, C.Y., Chan, T.F., Daum, C.G*., et al*. (2019). Construction and comparison of three reference-quality genome assemblies for soybean. Plant J *100*, 1066-1082.

145. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A*., et al*. (2001). The sequence of the human genome. Science *291*, 1304-1351.

146. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet *101*, 5-22.

147. Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing *17*, 395-416.

148. Wang, B., Zhu, Y., Zhu, J., Liu, Z., Liu, H., Dong, X., Guo, J., Li, W., Chen, J., Gao, C*., et al*. (2018a). 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.

149. Wang, C., Yang, Y., Yuan, X., Xu, Q., Feng, Y., Yu, H., Wang, Y., and Wei, X. (2014a). Genome-wide association study of blast resistance in indica rice. BMC Plant Biol *14*, 311.

150. Wang, Q., Tian, F., Pan, Y., Buckler, E.S., and Zhang, Z. (2014b). A SUPER powerful method for genome wide association study. PLoS One *9*, e107684.

151. Wang, S., He, S., Yuan, F., and Zhu, X. (2017). Tagging SNP-set selection with maximum information based on linkage disequilibrium structure in genome-wide association studies. Bioinformatics *33*, 2078-2081.

152. Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R.R., Zhang, F*., et al*. (2018b). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature *557*, 43-49.

153. Wang, X., Jia, M.H., Ghai, P., Lee, F.N., and Jia, Y. (2015). Genome-Wide Association of Rice Blast Disease Resistance and Yield-Related Components of Rice. Mol Plant Microbe Interact *28*, 1383-1392.

154. Wang, X., Niu, Q.W., Teng, C., Li, C., Mu, J., Chua, N.H., and Zuo, J. (2009). Overexpression of PGA37/MYB118 and MYB115 promotes vegetative-to-embryonic transition in Arabidopsis. Cell Res *19*, 224-235.

155. Wang, Y., He, J., Yang, L., Wang, Y., Chen, W., Wan, S., Chu, P., and Guan, R. (2016). Fine mapping of a major locus controlling plant height using a high-density single-nucleotide polymorphism map in Brassica napus. Theor Appl Genet *129*, 1479-1491.

156. Wang, Y., Zhang, W.Z., Song, L.F., Zou, J.J., Su, Z., and Wu, W.H. (2008). Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in Arabidopsis. Plant Physiol *148*, 1201-1211.

157. Wei, W., Mesquita, A.C.O., Figueiro, A.A., Wu, X., Manjunatha, S., Wickland, D.P., Hudson, M.E., Juliatti, F.C., and Clough, S.J. (2017). Genome-wide association mapping of resistance to a Brazilian isolate of Sclerotinia sclerotiorum in soybean genotypes mostly from Brazil. BMC Genomics *18*, 849.

158. Westra, H.J., Martinez-Bonet, M., Onengut-Gumuscu, S., Lee, A., Luo, Y., Teslovich, N., Worthington, J., Martin, J., Huizinga, T., Klareskog, L.*, et al.* (2018). Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. Nat Genet *50*, 1366-1374.

159. Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

160. Wu, X., Heffelfinger, C., Zhao, H., and Dellaporta, S.L. (2019). Benchmarking variant identification tools for plant diversity discovery. BMC Genomics *20*, 701.

161. Wulff, B.B., and Moscou, M.J. (2014). Strategies for transferring resistance into wheat: from wide crosses to GM cassettes. Front Plant Sci *5*, 692.

162. Xiao, Y., Liu, H., Wu, L., Warburton, M., and Yan, J. (2017). Genome-wide Association Studies in Maize: Praise and Stargaze. Mol Plant *10*, 359-374.

163. Xie, W., Wang, G., Yuan, M., Yao, W., Lyu, K., Zhao, H., Yang, M., Li, P., Zhang, X., Yuan, J.*, et al.* (2015). Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. Proc Natl Acad Sci U S A *112*, E5411-5419.

164. Yang, N., Liu, J., Gao, Q., Gui, S., Chen, L., Yang, L., Huang, J., Deng, T., Luo, J., He, L.*, et al.* (2019). Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet *51*, 1052-1059.

165. Yano, K., Morinaka, Y., Wang, F., Huang, P., Takehara, S., Hirai, T., Ito, A., Koketsu, E., Kawamura, M., Kotake, K.*, et al.* (2019). GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. Proc Natl Acad Sci U S A *116*, 21262-21267.

166. Yu, J., Golicz, A.A., Lu, K., Dossa, K., Zhang, Y., Chen, J., Wang, L., You, J., Fan, D., Edwards, D.*, et al.* (2019). Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. Plant Biotechnol J *17*, 881-892.

167. Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X.*, et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science *296*, 79-92.

168. Zhang, C., Dong, S.S., Xu, J.Y., He, W.M., and Yang, T.L. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics *35*, 1786-1788.

169. Zhang, S., Yu, H., Wang, K., Zheng, Z., Liu, L., Xu, M., Jiao, Z., Li, R., Liu, X., Li, J.*, et al.* (2018). Detection of major loci associated with the variation of 18 important agronomic traits between Solanum pimpinellifolium and cultivated tomatoes. Plant J.

170. Zhao, H., Yao, W., Ouyang, Y., Yang, W., Wang, G., Lian, X., Xing, Y., Chen, L., and Xie, W. (2015). RiceVarMap: a comprehensive database of rice genomic variations. Nucleic Acids Res *43*, D1018-1022.

171. Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326-3328.

172. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet *9*, e1003264.

173. Zhou, X., and Huang, X. (2019). Genome-wide Association Studies in Rice: How to Solve the Low Power Problems? Mol Plant *12*, 10-12.

174. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat Genet *44*, 821-824.

175. Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y.*, et al*. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol *33*, 408-414.

176. Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C.*, et al*. (2018). Rewiring of the Fruit Metabolome in Tomato Breeding. Cell *172*, 249-261 e212.