Haplotype based Genome-Wide Association Studies in peanut.

Ziqi Han
1012570
Plant breeding and genetic resources.
January 2024

M.Sc. Thesis
Quantitative aspects of Plant Breeding Group
Report no.

WAGENINGEN UNIVERSITY

Supervisors: Chris Maliepaard
Yanlin Liao

DATA FOR INTERNAL USE ONLY
EXTERNAL USE OF DATA IS ONLY PERMITTED WITH CONSENT OF THE PROJECT LEADER

Table of contents

Abstract/Summary

This study focuses on cultivated peanuts, *Arachis hypogaea*. The genetic makeup of peanuts is complex due to its allotetraploid nature, which was derived from the hybridization of two wild diploids (*A. duranensis* and *A. ipaensis*). A limited number of elite cultivars has led to reduced genetic diversity, prompting the need for improved breeding strategies. To analyze the link between genetic and phenotypic information, analyzing Quantitative Trait Loci (QTL) can provide insights into the genetic factors underlying variation in complex traits. Advancements in genotyping methods, including the use of high-density arrays and techniques like genotyping by sequencing (GBS), have greatly eased the process of identifying QTLs associated with various traits in different crops. Genome-wide association studies (GWAS) is an important genomics methodology that can identify QTL associated with complex quantitative traits using natural population.

In peanuts, most of research employs SNP-based GWAS to identify candidate regions related to traits of interest, such as yield-related traits, seed-related traits, and disease resistance. However, the issue of "missing heritability" is commonly observed in SNP-based GWAS, which means despite the successful identification of genetic variations associated with various traits and diseases through GWAS, the contribution of identified genetic variations cannot be fully explained or is limited. This issue may be caused by various factors, building haploids from adjacent SNPS is one approach to address it.

The use of haplotypes, defined from linked SNPs, has emerged as a methodological variant for identifying genomic regions from GWAS. In wheat and other crops, haplotype-based GWAS have already been applied and yielded promising results. But only limited research on haplotype-based GWAS exists for peanuts. This study initially aims to establish a general pipeline for haplotype-based GWAS in peanuts, followed by the investigation and selection of suitable packages for implementation. Finally, we validated the pipeline of haplotype-based GWAS on peanuts using the phenotypic and genotypic data from a diverse panel of cultivated peanuts. Four phenotypes associated with agronomic traits were studied. By investigating general methods for conducting haplotype-based GWAS, this research has the potential to contribute to future work in peanut genetics and breeding.

Introduction

Peanut

Peanut (*Arachis hypogaea*) is an annual legume crop, commonly referred to as groundnut (Asakura and Kitahora, 2013). It is a crucial crop with diverse applications worldwide, including food, oil, and seeds. It holds a significant position as one of the major oilseed and food crops. The market demand for peanuts is steadily growing, with unique requirements in different regions. A comprehensive understanding of the peanut genome is essential for breeders to efficiently optimize specific regions based on market demands (Arya *et al.,* 2016).

The cultivated peanut is taxonomically classified within the *Arachis* genus based on its morphological characteristics and cross-compatibility relationships with other species (Stalker *et al.,* 2016). *Arachis* species are predominantly found in tropical and subtropical regions, suggesting a probable origin in tropical wetland areas before adapting to survive in arid environments (Simpson *et al.,* 2001). The most plausible place of origin for cultivated peanuts is identified as northern Argentina and southern Bolivia (Stalker *et al.,* 2016).

The morphological features of peanuts, as per Ikisan.com sourced from Ikisan Agri-Informatics & Services Division of Nagarjuna Fertilizers and Chemicals Ltd (NFCL) (2021), include a well-defined main stem with a variable number of lateral branches. Groundnuts exhibit two recognized growth habits: prostrate and erect. The prostrate form is characterized by an upright and prominent main stem with procumbent or decumbent lateral branches. In contrast, the main axis loses its distinction from the laterals in erect types. (Source : Ikisan.com, https://www.ikisan.com/tn-groundnut-morphology.html).

During the pre-flowering phase, the arrangement of vegetative branches and inflorescences in the leaf axils on both the main axis and branches differs between the two primary botanical sections of *A. hypogaea* (Ikisan Agri-Informatics & Services Division of NFCL, 2021). Consequently, two flowering patterns exist in cultivated peanuts: a sequential pattern and an alternative pattern. In cultivated peanuts, the main branch (axis) is denoted as 'n', with subsequent branches termed 'n+1', 'n+2', and 'n+3'. Across all species forms, primary vegetative branches (n+1) emerge on the axis of cotyledons and at various higher nodes on the main axis. In sequential types, inflorescences develop at the second and several subsequent nodes of primary branches. The first node on a branch may produce a secondary branch (n+2), but often it bears an inflorescence, initiating flower development shortly after the n+1 branch. In alternative types, the first two nodes of the n+1 branch typically yield vegetative branches (n+2), followed by two nodes with inflorescences, and the pattern repeats with vegetative

branches and so on. This sequential pattern is mirrored in the n+2 branches(Ikisan Agri-Informatics & Services Division of NFCL, 2021). Figure 1 shows two plants with these two flowering patterns in (a) and (b).

Nowadays, the cultivation of peanuts has been extended from South America to over a hundred countries worldwide (Jati *et al.,* 2013). The steadily growing consumption of peanuts is due to its unique flavor, high nutritional value, high oil and protein level seeds and versatile uses (Allen *et al.,* 2013). Peanut oil is considered a premier frying oil due to its stability at high temperature and a high smoke point compared to other edible oils, resulting in excellent sensory properties and extended fry life (List *et al.,* 2016). It also contains sterols, such as β-sitosterol, known for inhibiting cancer growth and providing protection against colon, prostate, and breast cancer (Sanders *et al.,* 2003). Additionally, Rachaputi(2016) also mentioned that peanut and peanut butter in diet was associated with a 21% reduction in the risk of cardiovascular disease, whereas a low-fat diet resulted in only a 12% decrease in risk.

Peanut Genome

Polyploids can be classified into two types based on their different origins: allopolyploids and autopolyploid (Soltis *et al.,* 2000). Allopolyploids typically result from the hybridization of two different species, and they often exhibit bivalent pairings of chromosomes during meiosis, where more similar chromosomes are more likely to pair with each other (Xu *et al.,* 2013). Autopolyploid, on the other hand, arise from combinations of different genomes within the same species, and during meiosis, chromosomes can pair among more than two homologous copies (Xu *et al.,* 2013).

A cultivated peanut is a self-pollinated species and an allotetraploid (AABB, 2n = 4x = 40), resulting from a singular hybridization event between two wild diploids (Bertioli *et al.,* 2011). Kochert (1996) indicates that *Arachis duranensis* and *Arachis ipaensis* are the donor of A and B sub genome respectively. As reported by Samoluk et al. (2015), the genome sizes of the *A. duranensis* and *A. ipaensis* are approximately 1.25 Gb and 1.56 Gb, respectively. The sum of their genome sizes is close to the total genome size of *A. hypogaea*, which is approximately 2.8 Gb, suggesting that significant changes in genome size have not occurred since the polyploidization event (Temsch and Greilhuber, 2000, Lu *et al.,* 2018).

In peanut, the utilization of a limited number of elite cultivars has led to a narrow genetic basis and a diminished level of germplasm polymorphism (Fonceka *et al.,* 2009). When considering the framework of a breeding program for peanut enhancement, three foundational components emerge, including germplasm management, research priority areas, and breeding strategies (Coulibaly *et al.,* 2022). So, the primary breeding objectives of domesticated peanuts involve elevating the genetic potential of both

qualitative and quantitative traits, while concurrently enhancing genetic diversity and refining trait quality.

Haplotype

According to the definition from National Human Genome Research Institute, A haplotype refers to a clustered arrangement of genomic variants along a single chromosome that are commonly passed down together. It usually represents a distinct combination of variants located in proximity on a chromosome (https://www.genome.gov/genetics-glossary/haplotype).

Haplotype can be built based on the phased genotyping marker; genotype phasing method varies based on different type of polyploid. Package FitTetra 2.0 (Zych *et al.,* 2019) can be used for genotype calling for tetraploids, and package Beagle 5.4 (Browning *et al.,* 2021) can be used for allotetraploid genotyping. PolyHaplotyper is a haplotyping tool for polyploid species genetic analysis, based on bi-allelic markers such as SNPs (Voorrips *et al*., 2022). The package Beagle stands as an accurate phasing algorithm that is designed to efficiently handle large-scale genetic datasets (Browning *et al.*, 2007).

The GHap software package is used for haplotype extraction (Utsunomiya et al., 2020). Specifically designed for haplotype construction, it employs user-defined haplotype blocks to identify diverse haplotype alleles within the dataset. The package evaluates sample haplotype allele genotypes by considering the haplotype allele dosage (i.e., 0, 1, or 2 copies in a diploid). The resulting output is not only compatible with analyses involving multi-allelic markers but is also conveniently structured for integration into existing pipelines designed for bi-allelic markers. Originally introduced by Utsunomiya et al. (2016), the GHap software package streamlines haplotype construction from phased marker data, providing a robust foundation for subsequent analyses.

The precision of haplotype construction may vary depending on various factors such as sample size, SNP count, allele frequency, proportion of missing data, genotyping error rate, and the extent of linkage disequilibrium among these SNPs (Kirk and Cardon, 2002). Simulating datasets with varying SNP densities to determine the optimal number of SNPs within a haplotype should be considered (Zhang, 2004).

Genome-wide association study

Genome-wide association studies originated in human genetics, to detect the association between common genetic variants and the risk of human disease (Hirschhorn and Mark, 2005; Smith *et al.,* 2019). Over time, the application of GWAS has become widespread, extending beyond human genetics to include model organisms

in both the animal and plant kingdoms, as well as non-model systems (Korte and Farlow, 2013). For instance, researchers routinely utilize GWAS to identify specific genetic loci and underlying genetic structures associated with phenotypes determining various agronomically important traits in crops (Korte and Farlow, 2013). As advancements in statistical methodologies facilitate GWAS, opportunities arise to identify associations between phenotypic traits and specific genetic regions (Uffelmann *et al.,* 2021). Examples of GWAS applications include studying flowering time and grain yield traits in rice germplasm (Huang *et al.,* 2012), exploring agronomic and morphologic traits in barley cultivars (Wang *et al.,* 2012), and investigating disease resistance regions in wheat through GWAS (Malosetti *et al.,* 2020).

Improving genetic tools for peanuts, particularly through SNP-based GWAS, can enhance our knowledge of peanut genetic structures and identify specific candidate gene regions. SNP stands as one of the most prevalent genetic variations that can be regarded as the markers in genetic studies (Bush and Moore, 2012). This method holds the potential to facilitate crop improvement and address genetic barriers, contributing to the sustainable development of peanut cultivation. Although single-marker-based GWAS successfully identifies genetic variations associated with diverse traits and diseases, a frequently observed issue in SNP-based GWAS is 'missing heritability,' which means the contribution of identified genetic variations cannot be fully explained or is limited (Sehgal *et al.,* 2020). Several factors may contribute to this issue, and one strategy to address it involves constructing haplotypes from adjacent SNPs (Sehgal *et al.,* 2020).

Nowadays, GWAS has increasingly been employed to investigate the genetic foundations of significant characteristics in peanuts (Wang *et al.,* 2019). Despite a bunch of research employing SNP-based GWAS to identify candidate SNPs related to traits such as yield-related traits (Wang *et al.,* 2019), growth habit-related traits (Li *et al.,* 2022), oil content (Wang *et al.,* 2018), sting nematode resistance (Ravelombola *et al.,* 2022), and lead spot resistance (Zhang *et al.,* 2020). While several research groups have already investigated breeding traits in peanut using haplotype-based GWAS, recent studies have shown that haplotype-based GWAS can provide valuable supplementary information in diploid species like maize (Wang *et al.,* 2018, Hang *et al.,* 2020, Maldonado *et al.,* 2019). However, haplotype-based GWAS is still underutilized in peanut research. Establishing a general haplotype based GWAS pipeline for peanuts would simplify and enhance the application of this method for researchers.

The primary aim of this project is to develop a general haplotype-based GWAS pipeline for peanuts. The raw genotype data and phenotype data are provided by Henan Academy of Agricultural Sciences (HAAS). The first step involves an investigation into suitable packages for haplotyping peanuts, elucidating the preprocessing steps required

for raw data, and obtaining final haplotype information. Subsequently, GWAS will be performed using these haplotypes. Additionally, a comparative analysis with the HAAS SNP-based GWAS, conducted on the same peanut population genome, will be undertaken. This comparative approach aims to assess whether haplotype based GWAS can provide supplementary insights, offering an extra layer of information.

Materials & Method

Plant material

The peanut diversity panel comprises 353 accessions of cultivated tetraploid A. *hypogaea.* phenotypic and genotyping data were collected by the Institute of Crop Molecular Breeding at the Henan Academy of Agricultural Sciences.

Phenotyping traits

The phenotypic traits under consideration encompass binary and discrete attributes, including flowering pattern (alternate or sequential), inner integument color (yellow or white), growth habit (erect or prostrate), and the total number of branches. Examples of these phenotype traits are depicted in Figure 1. The seeds of each line were sown using a randomized complete block design with two replicates within a single environment. The phenotype data is provided by HAAS.

a.

b.

c.

d.

e.

f.

Figure 1 : Phenotypic Traits. Illustrative examples of distinct phenotypic traits are presented:
a) alternate flowering pattern: The green triangles indicate the vegetative branches, the red triangles the nodes with inflorescences. b) sequential flowering pattern. c) yellow seed (inner integument). d) white seed. e) prostrate growth habit f) erect growth habit.

Genotyping material

Whole-genome resequencing was carried out across various *Arachis* species and then aligned against the genome of the peanut cultivar Tifrunner (Li, 2018). Paired end DNA libraries were formed with around 300 bp inserts, and subsequently, sequencing was conducted using the Illumina HiSeq Xten platform (Illumina, Inc., San Diego, CA, USA) with a PE151 configuration. After undergoing quality checks and filtering, the superior quality reads were aligned to the genome of cultivated peanut (*Arachis hypogaea* cv. Tifrunner version 1) using the minimap2 (v2.10) software. SNP and INDEL calling were performed with the Genome Analysis Toolkit. After applying quality control, a total of 864,179 SNPs and 71,052 InDels were identified. The raw genotyping material was provided by HAAS.

Genotyping

The primary steps of haplotype based GWAS were designed as follows: First, we did the genotyping and constructed haplotypes, followed by a construction of a kinship

matrix. Finally, we performed GWAS using a mixed linear model, accounting for kinship.

Initially, we employed both updog (Gerard *et al.,* 2018) and beagle 5.4 (Browning *et al.,* 2021) for genotyping from read counts, treating genotypes as diploid. Beagle 5.4 was also utilized for phasing and imputation. Subsequently, we conducted a comparative analysis between the outcomes of these two tools.

Using the R package vcfR, we extracted the 'gt' field from the raw VCF file, facilitating the retrieval of total read depth (DP) and reference read counts (AD) per SNP. These values were then utilized to construct 'refmat' and 'sizemat.' Following data processing, the 'multidog' function in the R package 'updog' was employed to obtain genotyping results. The script for updog application can be checked in the appendix. Simultaneously, the 'beagle 5.4' package was utilized for genotyping, imputation, and phasing, executed through the Java script 'java -jar beagle.22Jul22.46e.jar gt=s353.gwas.recode93.vcf out=out.gt' in Linux. The resultant beagle outcomes were juxtaposed with those of updog.

To assess the quality of genotyping data for each SNP, we implemented criteria to remove SNPs with low call rates (SNPs with Minor Allele Frequency, MAF < 0.05) and high missing data (Missing data > 0.1). The result of beagle was then used to impute the missing data in the result of updog. Given the uncertain order of individual parents of heterozygotes in updog, this information was filled using beagle's results.

Haplotyping

To derive haplotype results, we utilized the R-package GHap, employing essential input files draw information from both Beagle and Updog results. According to the requirement, the one of input files should be devoid of missing data and accurately represent the phased chromosome alleles. To achieve this, we utilize Beagle results to impute missing data in the Updog-derived dataset. The GHap package's ghap.blockgen() function is subsequently employed to delineate haploblocks, and the ghap.haplotyping() function is utilized to generate a matrix of haplotype genotypes. This sequential process ensures the creation of a robust and comprehensive dataset suitable for downstream haplotype analysis.

GWAS

GWAS was conducted to assess the associations between haplotypes and phenotypic traits, employing logistic regression models and accounting for potential population structure and relatedness. Multiple testing correction methods were applied to ensure the robustness of the results. In this case, due to the limited time, the covariate was only

set as kinship matrix. The kinship matrix of hapallele was built in r package GHap, using ghap.kinship() function.

After getting the kinship matrix, a GWAS was performed utilizing a mixed linear model, incorporating kinship matrix as covariate. To enhance the reliability of findings, a comprehensive approach that incorporates both TASSEL and GHap software platforms was embraced. Visualizing the outcomes was done through Manhattan plots and QQ plots. The script of GHap GWAS is shown in the appendix. Due to the large dataset, we perform TASSEL 5 in Linux using command line "./run_pipeline.pl -Xms512m -Xmx10g -fork1 -plink -ped res_pop.ped -map res_pop.map -sortPositions -fork2 -r phenotype.txt -fork3 -t traits.txt -fork4 -k res_pop_kinship.txt -combine5 -input1 -input2 -intersect -combine6 -input5 -input4 -mlm -mlmVarCompEst P3D -mlmCompressionLevel Optimum -mlmOutputFile gwas_result".

The result of this haplotype-based GWAS was compared with that of the existing SNP-based GWAS which uses the same genotyping dataset. We compared the visualization of associations using QQ plot and Manhattan plot. QQ plot was used to compare statistical significance, overlay the QQ plot from both haplotypes based GWAS and SNP-based GWAS to visually compare the distribution of observed p-value. Deviations from the expected distribution in one direction (above the diagonal) can indicate a higher number of significant associations compared to the other method. Manhattan plot was used to compare the candidate genetic region number to explore whether there are unique associations detected by either method.

Result

Genotyping result

The raw genetic variations within the peanut gene sequence are stored in the 's353.gwas.recode93.vcf' file, capturing data from 353 samples, comprising 864,179 SNPs, and 71,052 indels. The total number of sites under analysis was 935,231. Initially, we employed both updog (Gerard *et al.,* 2018) and beagle5.4 (Browning *et al.,* 2021) for genotyping based on read counts, treating genotypes as diploid, followed by a comparative analysis of the results.

Compared to the results obtained from the Updog package, we noted occasional misclassification of some homozygous genotypes as heterozygous by Beagle, which was an unexpected outcome. In contrast, the genotyping accuracy demonstrated by the updog outcomes appears superior relative to those obtained with Beagle. However, it's important to note that Beagle performed imputation, eliminating missing data, and also provided additional phasing information for heterozygous individuals.

We assessed the quality of genotyping data for each SNP and eliminated SNPs with MAF < 0.05 and Missing data > 0.1 in the updog results. Following this quality filter, 578,088 samples remained. Figure 1 illustrates the dynamic fluctuations in SNP counts across chromosomes and demonstrates the impact of filtering.

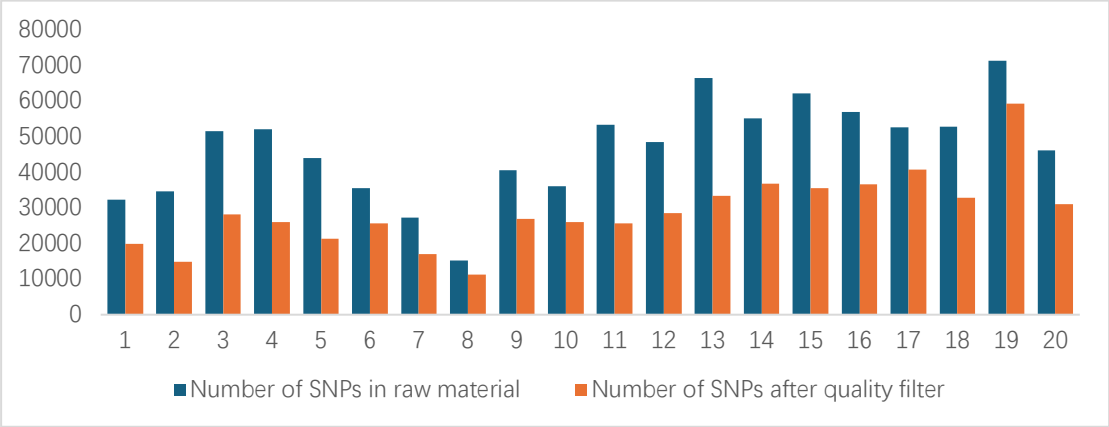


Figure 2. Portray the dynamic alteration in the number of SNPs per chromosome, encompassing both pre- and post-quality filtering stages. The x-axis delineates the chromosomes, while the y-axis denotes the SNP count.

We extract the Beagle results specifically for the subset of 578,088 samples from the comprehensive Beagle dataset, maintaining consistency with the filtering criteria applied to the updog results. Following this, in R, we rectify the inversion of 'alt' and

'ref' in the Beagle results by implementing necessary adjustments. Finally, we utilize the beagle results to impute missing data in the updog results. Given the uncertain order of individual parents of heterozygotes in updog, we incorporate this information from beagle's results.

Haplotype result

We obtained a total of 115,609 haploblocks, each containing 5 SNPs per block, using the GHap package. Subsequently, from these haploblocks, we derived a total of 763,207 haplotypes. The results of haplotyping revealed that the number of haplotypes per block varies from 2 to 19, with the majority falling within the range of 2 to 10. Figure 3 offers a concise overview of haplotype distribution across the genome, providing insights into the chromosomal variation and diversity captured by the analysis. Additionally, Table 1 presents the number of haploblocks and haplotypes per chromosome (Appendix).

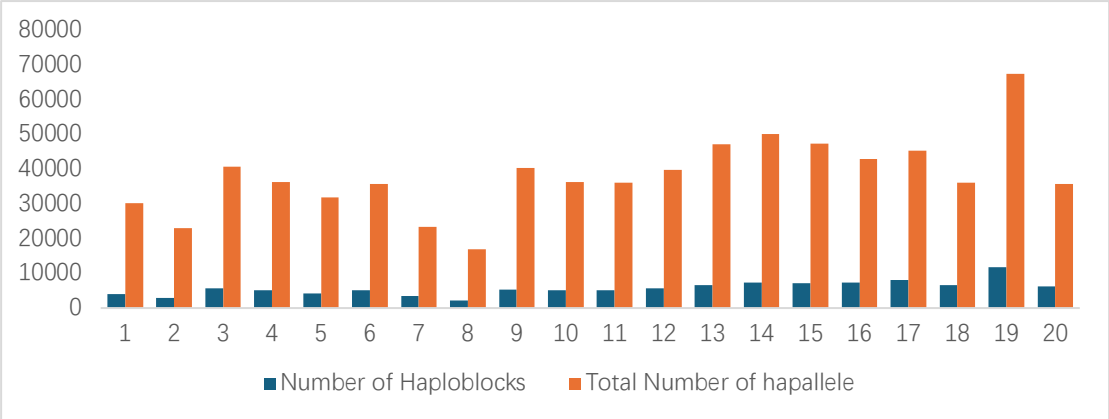


Figure 3: Haplotypes. Number of haploblocks and haplotypes per chromosome. Each haploblock comprises 5 SNPs. The x-axis represents individual chromosomes, while the y-axis signifies the haploblock count.

Haplotype-based GWAS

The kinship matrix was obtained using the GHap package and used as input for TASSEL. The heatmap of this kinship matrix is depicted in Figure 4 (Appendix). While GHap can be directly utilized for haplotype-based GWAS, we explored three methods . Method 1 involved GWAS without consider the effect of kinship and permanent environmental effect (repeated measurement, rep), Method 2 included setting kinship as a covariate but without rep, and the last method comprised performing GWAS only with repeated measurements.

Due to formulation limitations, we could only obtain results when performing GWAS

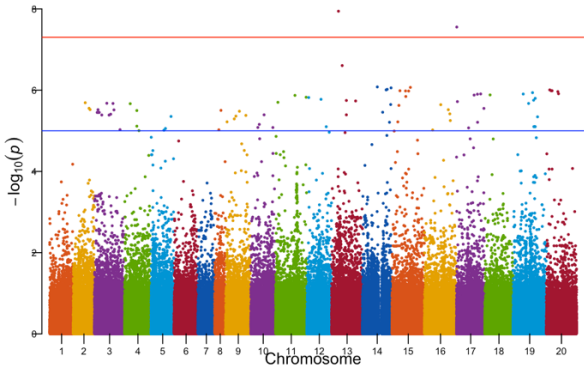
without considering kinship. In the quest for an alternative package that accounts for kinship, we turned to the TASSEL package in Linux. The execution of GWAS using the mixed linear model took more than one week to yield results for the four phenotypes. Figures 3b) and 3d) present the Manhattan plots and QQ plots for the SNP-based GWAS of phenotype traits, including the total number of branches and growth habits from HAAS. Notably, based on the QQ plot and Manhattan plot, we observed suboptimal GWAS results for flowering pattern and inner integument color.

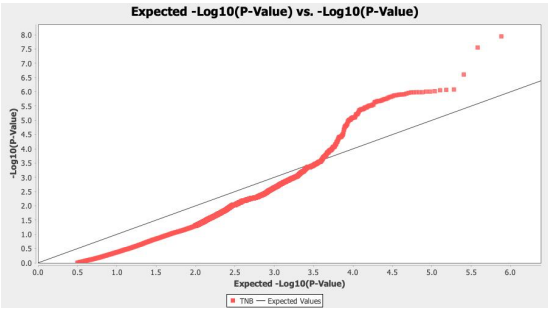
Due to formulation limitations, we could only obtain results when performing GWAS without considering kinship. Seeking an alternative package that accounts for kinship, we employed the TASSEL package in Linux. The execution of GWAS using the mixed linear model took more than one week to obtain results for the four phenotypes.

According to the haplotype-based GWAS visualization, in Figure 3a), it is evident that chromosomes 13 and 17 contains potential candidate regions with a significant association with the total number of branches. Figure 3c) shows that chromosome 3 may contains the most associated intervals with the growth pattern. We compared our results with the HAAS group study, which utilized the same genotype and phenotype material to perform SNP-based GWAS. Figures 3b) and 3d) present the Manhattan plots and QQ plots for the SNP-based GWAS of phenotype traits, including the total number of branches and growth habits from HAAS. In the SNP-based GWAS Manhattan plot, the genomic regions which associate with total number of branches were identified on chromosomes 2, 9, 12, 15 and 16. For the growth habit, the associated genomic regions were situated chromosomes 3 and 15. We found that haplotype-based result shows some different genomic regions which may associate with these two phenotype traits.

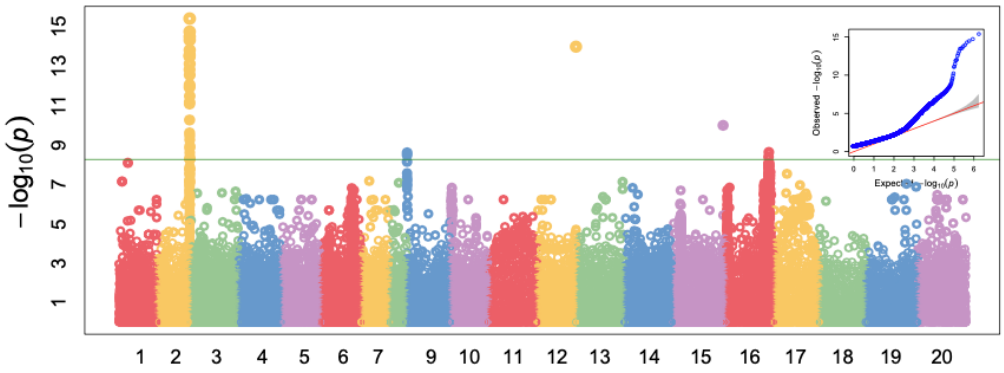
Notably, based on the QQ plot and Manhattan plot in Figure 3e) and 3f), we observed suboptimal GWAS results for flowering pattern and inner integument color. In both QQ plots, majority of points deviate above the diagonal line, indicating that the observed P-values for most sites exceeds the expected values. This observation implies a significant association for numerous loci with the flowering pattern and inner integument color, a trend that appears incongruent with biological logic.

a

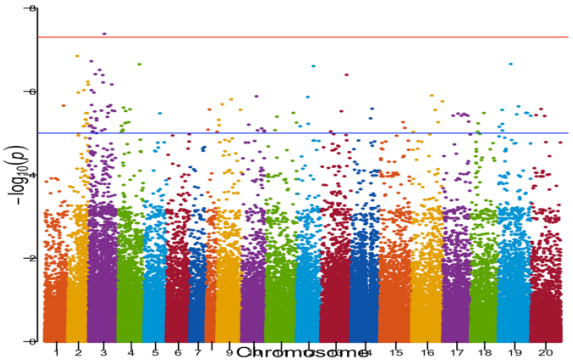


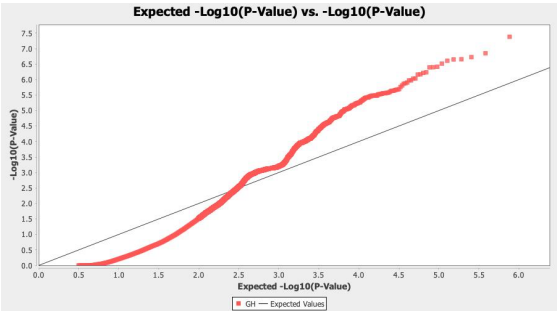


b

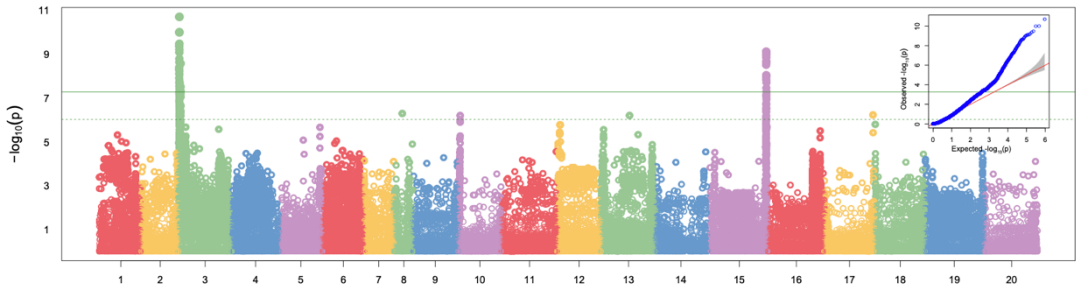

$-\log_{10}(p)$
1 3 5 7 9 11 13 15
1 2 3 4 5 6 7 9 10 11 12 13 14 15 16 17 18 19 20
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

c

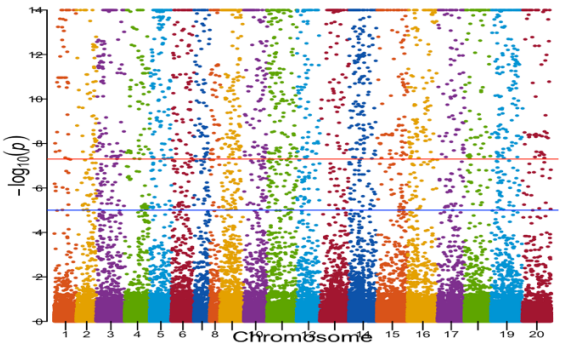

$-\log_{10}(p)$
1 2 3 4 5 6 7 9 10 11 12 13 14 15 16 17 18 19 20
Chromosome

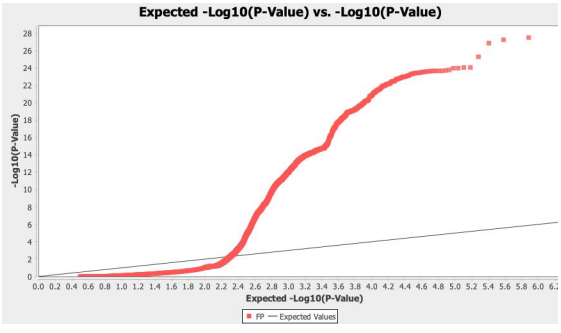

Expected -Log10(P-Value) vs. -Log10(P-Value)
-Log10(P-Value)
Expected -Log10(P-Value)
GH — Expected Values

d


$-\log_{10}(p)$
1 3 5 7 9 11
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

e


$-\log_{10}(p)$
1 2 3 4 5 6 8 15 16 17 19 20
Chromosome


Expected -Log10(P-Value) vs. -Log10(P-Value)
-Log10(P-Value)
Expected -Log10(P-Value)
FP — Expected Values

f

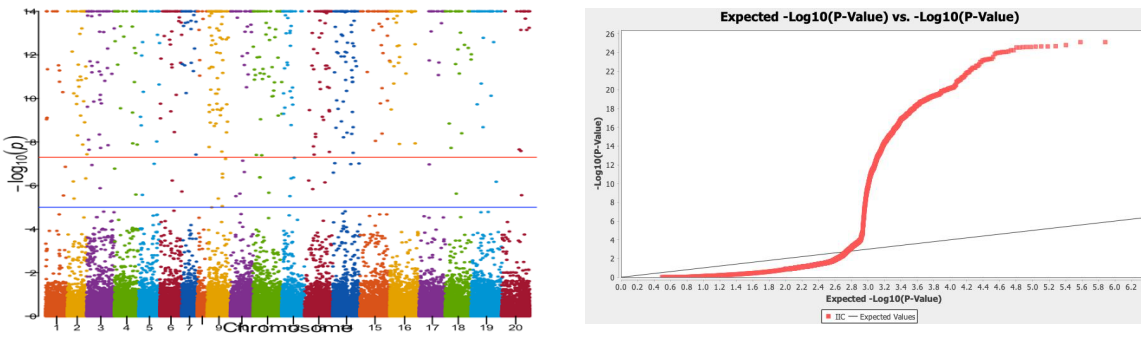


Figure 3: Haplotype Results and Peanut SNP-based GWAS result from HAAS. a) GWAS Manhattan plot and quantile-quantile QQ plot for total number of branches; b) Manhattan plots and QQ plots for the SNP-based GWAS of total number of branches from HAAS; c) Manhattan plot and quantile-quantile QQ plot for growth habit; d) Manhattan plots and QQ plots for the SNP-based GWAS of growth habit from HAAS; e) Manhattan plot and quantile-quantile QQ plot for flowering pattern; f) GWAS Manhattan plot and quantile-quantile QQ plot for inner integument color. In the haplotype-based GWAS Manhattan plot a), c), e) and f), the blue line is the threshold for suggestiveness of a QTL, the red line denotes the genome wide significance threshold. Suggestiveness line defaults to -log10(1e-5) and genome-wide significance line defaults to -log10(5e-8). In the SNP-based GWAS Manhattan plot b) and d), the green horizontal line indicates the threshold for significant association ($p<0.05$) after the Bonferroni correction.

Discussion

The outcomes of the haplotype-based GWAS for two phenotypic traits, flowering pattern, and inner integument color, did not align with our initial expectations. The QQ plot revealed a considerable number of false positives, indicating potential issues in the analysis. Several factors could contribute to this unexpected result, involving insufficient population structure analysis, phenotype data and limitation of the method.

It is important to acknowledge certain limitations in our study. Due to time constraints, in the GWAS mixed linear model, we focused only on the impact of kinship matrices, without considering population structure analysis. Literature review suggests the importance of addressing extra population structure analysis, for which both the STRUCTURE (Porras-Hurtado *et al.,* 2013) and TASSEL packages prove valuable in calculating suitable K values and conducting Principal Component Analysis (PCA) which can also be utilized as covariates in the GWAS mixed linear model.

Another consideration is the absence of replicated phenotype data and measurements taken under different environmental conditions. The limited input of phenotype data may impact the results, as environmental effects are sometimes considerable and can mask genotypic effects (Zaitlen and Kraft, 2012). Additionally, during the arrangement of genotyping results, we employed less accurate beagle results to impute the updog results, potentially introducing a negative impact on the overall outcome.

In regard to the selection of software and package tools throughout the entire pipeline, additional experimentation and comparison are needed. There are still many other packages that have not been used in this study. Therefore, there is still the possibility of finding packages that are more suitable for the entire haplotype-based GWAS, which is worth further exploration and experimentation. Additionally, new software packages and tools are continually being published in the academic community. Hence, for the software packages chosen for each step in the entire workflow, it is worthwhile for us to continue to pay attention to and explore.

Furthermore, during the construction of Haploblocks using GHap, we set the threshold at 5 SNPs per block without considering other SNP counts. It's worth noting that the number of SNPs is a critical factor that can impact the precision of haplotype construction (Kirk and Cardon, 2002; Zhang, 2004). In this study, to explore this, we created different haploblocks with 4 and 6 SNPs per chromosome. Unfortunately, due to time constraints, we were unable to test these alternatives. Future studies may benefit from exploring varying SNP count thresholds to enhance the accuracy of haplotyping construction.

Conclusion

The overall pipeline for haplotype-based GWAS involves a systematic series of steps, encompassing phenotype and genotype data collection, genotyping, genotype quality filtering, haplotyping, kinship matrix construction, GWAS execution, and result visualization. Throughout the study, a comprehensive set of software tools and packages can be strategically employed. For genotyping, the utilization of the updog package is recommended, supplemented by phased information and imputation obtained through Beagle 5.4. Setting genotyping data quality filter thresholds at MAF ≤ 0.05 and missing data > 0.1 is advised. GHap package can be used for haplotyping, although alternatives like Haploview can also be explored. The different SNP count setting per haploblock can be applied. The analysis of kinship matrices can be efficiently performed by GHap, with additional options including Haploview, TASSEL, and the R-package GAPIT3 (Wang *et al.,* 2021). PCA can also be consider using package STRUCTURE and TASSEL. Subsequently, GWAS can be executed using Tassel 5, GHap, and GAPIT3 to identify genomic regions associated with four key phenotypic traits. To ensure data format compatibility, file format conversion can be accomplished using Plink and Tassel 5.

In conclusion, this study proposes a powerful haplotype-based GWAS pipeline and suggests the need to carefully consider population structure effects and improve the accuracy of haplotyping construction, and Continue to explore more suitable packages, providing avenues for future research. Further research should further optimize parameter selection and explore alternative methods to improve the comprehensiveness of genetic association studies.

References

Allen, L. H. (2013, January 1). Legumes (B. Caballero, Ed.). Retrieved November 9, 2021, from ScienceDirect website: https://www.sciencedirect.com/science/article/pii/B9780123750839001707

Arya, S. S., Salve, A. R., & Chauhan, S. (2016). Peanuts as functional food: a review. Journal of food science and technology, 53(1), 31–41. https://doi.org/10.1007/s13197-015-2007-9

Asakura, H., & Kitahora, T. (2013, January 1). Chapter 3 - Antioxidants in Inflammatory Bowel Disease, Ulcerative Colitis, and Crohn Disease (R. R. Watson & V. R. Preedy, Eds.). Retrieved January 22, 2024, from ScienceDirect website: https://www.sciencedirect.com/science/article/pii/B9780123971548000129

Bertioli, D. J., Seijo, G., Freitas, F. O., Valls, J. F. M., Leal-Bertioli, S. C. M., & Moretzsohn, M. C. (2011). An overview of peanut and its wild relatives. *Plant Genetic Resources*, *9*(01), 134–149. https://doi.org/10.1017/s1479262110000444

Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. American Journal of Human Genetics, 108(10), 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005

Browning, S. R., & Browning, B. L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLOS Computational Biology 8(12): e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Coulibaly, M., Guillaume Bodjrenou, Félicien Akohoue, Eric Etchikinto Agoyi, Francisco, Chaldia Oa Agossou,…Achigan-Dako, E. G. (2022). Profiling Cultivars Development in Kersting's Groundnut [Macrotyloma geocarpum (Harms) Maréchal and Baudet] for Improved Yield, Higher Nutrient Content, and Adaptation to Current and Future Climates. *Frontiers in Sustainable Food Systems*, *5*. https://doi.org/10.3389/fsufs.2021.759575

Fonceka, D., Téou Hodo-Abalo, Ronan Rivallan, Faye, I., Mbaye Ndoye Sall, Ndoye, O., … Rami, J.-F. (2009). *Genetic mapping of wild introgressions into cultivated peanut: a way toward enlarging the genetic basis of a recent allotetraploid*. *9*(1). https://doi.org/10.1186/1471-2229-9-103

Gawenda, I., Thorwarth, P., Torsten Günther, Ordon, F., & Schmid, K. (2015). Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. Plant Breeding, 134(1), 28–39. https://doi.org/10.1111/pbr.12237

Gerard, D., Ferrão, L. F. V., Garcia, A. A. F., & Stephens, M. (2018). Genotyping Polyploids from Messy Sequencing Data. Genetics, 210(3), 789–807. https://doi.org/10.1534/genetics.118.301468

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, *6*(2), 95–108. https://doi.org/10.1038/nrg1521

Ikisan Agri-Informatics & Services Division of Nagarjuna Fertilizers and Chemicals Ltd (NFCL). (2021). Groundnut Morphology. Ikisan.com. https://www.ikisan.com/tn-groundnut-morphology.html

Jati, I. R. A. P., Vadivel, V., & Biesalski, H. K. (2013, January 1). Chapter 31 - Antioxidant Activity of Anthocyanins in Common Legume Grains (R. R. Watson & V. R. Preedy, Eds.). Retrieved June 11, 2020, from ScienceDirect website: https://www.sciencedirect.com/science/article/pii/B9780123971548000075

Kirk, K. M., & Cardon, L. R. (2002). The impact of genotyping error on haplotype reconstruction and frequency estimation. *European Journal of Human Genetics*, *10*(10), 616–622. https://doi.org/10.1038/sj.ejhg.5200855

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, *9*(1), 29. https://doi.org/10.1186/1746-4811-9-29

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, L., Cui, S., Dang, P. et al. GWAS and bulked segregant analysis reveal the Loci controlling growth habit-related traits in cultivated Peanut (Arachis hypogaea L.). BMC Genomics 23, 403 (2022). https://doi.org/10.1186/s12864-022-08640-3

List, G. R. (2016, January 1). Chapter 15 - Processing and Food Uses of Peanut Oil and Protein (H. T. Stalker & R. F. Wilson, Eds.). Retrieved April 29, 2022, from ScienceDirect website: https://www.sciencedirect.com/science/article/pii/B9781630670382000150

Lu, Q., Li, H., Hong, Y., Zhang, G., Wen, S., Li, X., … Liang, X. (2018). Genome Sequencing and Analysis of the Peanut B-Genome Progenitor (Arachis ipaensis). *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00604

Maldonado, C., Mora, F., Scapim, C. A., & Coan, M. (2019). Genome-wide haplotype-based association analysis of key traits of plant lodging and architecture of maize identifies major determinants for leaf angle: hapLA4. *PLOS ONE*, *14*(3), e0212925. https://doi.org/10.1371/journal.pone.0212925

Malosetti, M., Zwep, L. B., Forrest, K., F.A. van Eeuwijk, & Dieters, M. J. (2020). Lessons from a GWAS study of a wheat pre-breeding program: pyramiding resistance alleles to Fusarium crown rot. *Theoretical and Applied Genetics*, *134*(3), 897–908. https://doi.org/10.1007/s00122-020-03740-8

National Human Genome Research Institute. (2019). Haplotype. Retrieved from Genome.gov website: https://www.genome.gov/genetics-glossary/haplotype

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., & Lareu, M. V. (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software. Frontiers in genetics, 4, 98. https://doi.org/10.3389/fgene.2013.00098

R.C.N Rachaputi, & Wright, G. (2016). Peanuts, Overview. *Elsevier EBooks*. https://doi.org/10.1016/b978-0-08-100596-5.00038-x

Sanders, T. H. (2003). GROUND NUT OIL. *Encyclopedia of Food Sciences and Nutrition*, *2*, 2967–2974. https://doi.org/10.1016/b0-12-227055-x/01353-5

Sanseverino, W., Roma, G., De Simone, M., Faino, L., Melito, S., Stupka, E., … Ercolano, M. R. (2009). PRGdb: a bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Research*, *38*(suppl_1), D814–D821. https://doi.org/10.1093/nar/gkp978

Sehgal, D., Mondal, S., Crespo-Herrera, L., Velu, G., Juliana, P., Huerta-Espino, J., Shrestha, S., Poland, J., Singh, R., & Dreisigacker, S. (2020). Haplotype-Based, Genome-Wide Association Study Reveals Stable Genomic Regions for Grain Yield in CIMMYT Spring Bread Wheat. Frontiers in Genetics, 11. https://doi.org/10.3389/fgene.2020.589490

Simpson, C. S., Krapovickas, A., & Montenegro, F. (2001). History of *Arachis* Including Evidence of *A. hypogaea* L. Progenitors. *Peanut Science*, *28*(2), 78–80. https://doi.org/10.3146/i0095-3679-28-2-7

Smith, C. J., Steinbrekera, B., & Dagle, J. M. (2019, January 1). Chapter 12 - Genetic Basis of Patent Ductus Arteriosus (R. K. Ohls, A. Maheshwari, & R. D. Christensen, Eds.). Retrieved December 13, 2023, from ScienceDirect website: https://www.sciencedirect.com/science/article/abs/pii/B9780323544009000126

Soltis, P. S., & Soltis, D. E. (2000). The role of genetic and genomic attributes in the success of polyploids. *Proceedings of the National Academy of Sciences*, *97*(13), 7051–7057. https://doi.org/10.1073/pnas.97.13.7051

Stalker, H. T., Tallury, S. P., Seijo, G. R., & Leal-Bertioli, S. C. (2016, January 1). Chapter 2 - Biology, Speciation, and Utilization of Peanut Species (H. T. Stalker & R. F. Wilson, Eds.). Retrieved January 22, 2024, from ScienceDirect website: https://www.sciencedirect.com/science/article/pii/B9781630670382000022

Temsch, E. M., & Greilhuber, J. (2000). Genome size variation in Arachis hypogaea and A. monticola re-evaluated. *Genome*, *43*(3), 449–451. https://doi.org/10.1139/g99-130

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., … Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-021-00056-9

Voorrips, R. E., & Tumino, G. (2022). PolyHaplotyper: haplotyping in polyploids based on bi-allelic marker dosage data. *BMC Bioinformatics*, *23*(1). https://doi.org/10.1186/s12859-022-04989-0

Waltram Ravelombola, Cason, J. M., Shyam Tallury, Manley, A., & Hanh Thi Pham. (2022). Genome-wide association study and genomic selection for sting nematode resistance in peanut using the USDA public data. *Journal of Crop Improvement*, *37*(2), 273–290. https://doi.org/10.1080/15427528.2022.2087127

Wang J, Yan C, Li Y, Li C, Zhao X, Yuan C, Sun Q, Shan S. GWAS Discovery Of Candidate Genes for Yield-Related Traits in Peanut and Support from Earlier QTL Mapping Studies. Genes (Basel). 2019 Oct 12;10(10):803. doi: 10.3390/genes10100803. PMID: 31614874; PMCID: PMC6826990.

Wang J., Zhang Z., GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction, Genomics, Proteomics & Bioinformatics (2021), doi: https://doi.org/10.1016/j.gpb.2021.08.005.

Wang, X., Xu, P., Liang, Y., Ren, Y., Li, S., Shi, Y., … Yuan, M. (2018). Genomic and Transcriptomic Analysis Identified Gene Clusters and Candidate Genes for Oil Content in Peanut (Arachis hypogaea L.). *Plant Molecular Biology Reporter*, *36*(3), 518–529. https://doi.org/10.1007/s11105-018-1088-9

Xu, F., Tong, C., Lyu, Y., Bo, W., Pang, X., & Wu, R. (2013). Allotetraploid and autotetraploid models of linkage analysis. *Briefings in Bioinformatics*, *16*(1), 32–38. https://doi.org/10.1093/bib/bbt075

Yuan, Y., Armin Scheben, Edwards, D., & Chan, T.-F. (2021). Toward haplotype studies in polyploid plants to assist breeding. *Molecular Plant*, *14*(12), 1969–1972. https://doi.org/10.1016/j.molp.2021.11.004

Yuri Tani Utsunomiya, Milanesi, M., Barbato, M., Taiti, A., Johann Sölkner, Paolo Ajmone-Marsan, & José Fernando Garcia. (2020). Unsupervised detection of ancestry tracks with the

GHaprpackage. *Methods in Ecology and Evolution*, *11*(11), 1448–1454. https://doi.org/10.1111/2041-210x.13467

Yuri Tani Utsunomiya, Milanesi, M., Taiti, A., Paolo Ajmone-Marsan, & José Fernando Garcia. (2016). GHap: an R package for genome-wide haplotyping. *Bioinformatics*, *32*(18), 2861–2862. https://doi.org/10.1093/bioinformatics/btw356

Zaitlen, N., & Kraft, P. (2012). Heritability in the genome-wide association era. Human Genetics, 131(10), 1655–1664. https://doi.org/10.1007/s00439-012-1199-6

Zhang, H., Chu, Y., Dang, P., Tang, Y., Jiang, T., Clevenger, J., … Chen, C. (2020). Identification of QTLs for resistance to leaf spots in cultivated peanut (Arachis hypogaea L.) through GWAS analysis. *Theoretical and Applied Genetics*, *133*(7), 2051–2061. https://doi.org/10.1007/s00122-020-03576-2

Zych, K., Gort, G., Maliepaard, C. A., Jansen, R. C., & Voorrips, R. E. (2019). FitTetra 2.0 - improved genotype calling for tetraploids with multiple population and parental data support. BMC bioinformatics, 20(1), 148. https://doi.org/10.1186/s12859-019-2703-y

Appendix

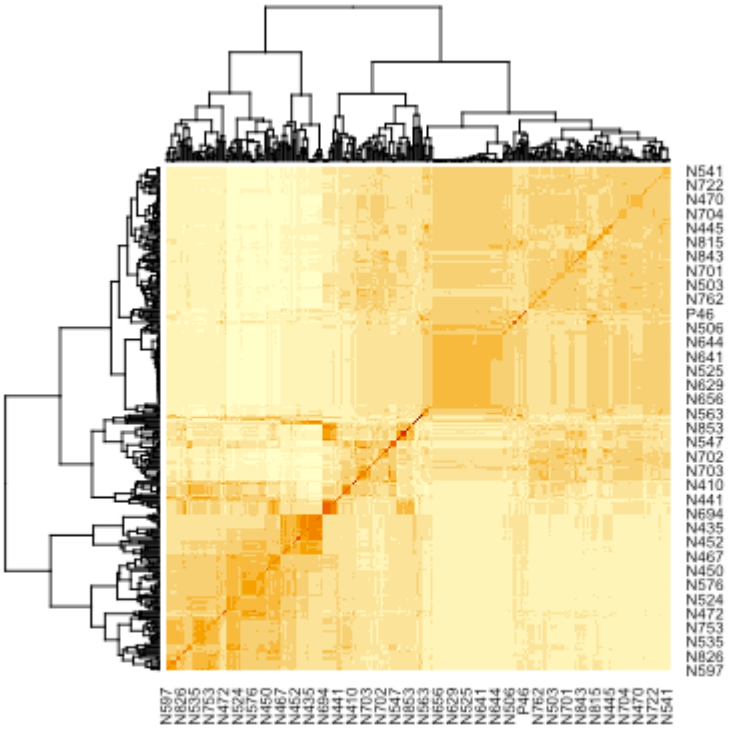


Figure 4. kinship heatmap. The kinship matrix, which is utilized to create the heatmap presented in Figure 4, was generated using the ghap.kinship() function in the R package GHap. This function facilitates the computation of HapAllele-based relationship matrices. Notably, in this case, each haploblock comprises 5 SNPs, and the resulting heatmap is based on hapallele. The associated color histogram illustrates the distribution of coefficients of coancestry, with intensified red hues emphasizing individuals with stronger relational ties.

Table 1. Presents the overall count of SNPs per chromosome both before and after the application of quality filters. Additionally, it provides information on the number of haploblocks per chromosome, as well as the total count of haplotypes per chromosome.

CHROME	Number of SNPs raw material	Number of SNPs quality filter	Number of Haploblocks	Total Number of hapallele (haplotype)
Arahy.01	32398	19980	3996	30138
Arahy.02	34688	14903	2980	22986
Arahy.03	51575	28115	5623	40825
Arahy.04	52067	26079	5215	36385
Arahy.05	43992	21392	4278	31921
Arahy.06	35628	25655	5131	35772
Arahy.07	27364	16974	3394	23388
Arahy.08	15234	11351	2270	16995
Arahy.09	40551	27003	5400	40373
Arahy.10	36092	26032	5206	36360
Arahy.11	53356	25575	5115	36034
Arahy.12	48438	28592	5718	39861
Arahy.13	66553	33347	6669	47117
Arahy.14	55211	36817	7363	50120
Arahy.15	62189	35624	7124	47442
Arahy.16	56870	36693	7338	42900
Arahy.17	52580	40819	8163	45273
Arahy.18	52800	32803	6560	36147
Arahy.19	71437	59283	11856	67413
Arahy.20	46208	31051	6210	35757
total	935231	578088	115609	763207

Use of scripts

1. Genotyping sample using r-package Updog script.

```
library(vcfR)
library(updog)
vcf <-read.vcfR("RAW/s353.gwas.recode93.vcf")
#open read counts using vcfR
fix_sample <- getFIX(vcf) %>% head
#Get elements from the fixed region of a VCF file

markernames_sample <- paste0(fix_sample[,1],'_',fix_sample[,2])
markernames_sample
#[1] "Arahy.01_90561" "Arahy.01_91340" "Arahy.01_93550" "Arahy.01_94281"
#[5] "Arahy.01_94568" "Arahy.01_94948"

gt_sample <- vcf@gt[1:6, 1:6]
rownames(gt_sample) <- markernames_sample
gt_sample1 <- gt_sample[which(gt_sample[,1] %in% 'GT:AD:DP:GQ:PGT:PID:PL'),]

gt_sample1
#             FORMAT                 JiNong99 ...
#Arahy.01_90561 "GT:AD:DP:GQ:PGT:PID:PL" "0/0:5,0:5:0:.:.:0,0,112" ...
#Arahy.01_94948 "GT:AD:DP:GQ:PGT:PID:PL"
"0/0:34,0:34:93:.:.:0,93,1395" ...

gt_sample2 <- gt_sample[which(!(gt_sample[,1] %in% 'GT:AD:DP:GQ:PGT:PID:PL')),]
gt_sample2
#             FORMAT          JiNong99
#Arahy.01_91340 "GT:AD:DP:GQ:PL" "0/0:7,0:7:21:0,21,269"
#Arahy.01_93550 "GT:AD:DP:GQ:PL" "0/0:12,0:12:33:0,33,495"
#Arahy.01_94281 "GT:AD:DP:GQ:PL" "0/0:22,0:22:66:0,66,822"
#Arahy.01_94568 "GT:AD:DP:GQ:PL" "0/0:17,0:17:51:0,51,591"

col_n_sample <- ncol(gt_sample)

#extract information: total number of read depth
dp1_sample <- do.call(cbind,lapply(2:col_n_sample,function(j){
  as.numeric(gsub('(.+)(:)(.+)(:)(.+)(:)(.+)(:)(.+)(:)(.+)(:)(.+)',
                  '\\5',gt_sample1[,j]))
}))
colnames(dp1_sample) <- colnames(gt_sample1)[2:col_n_sample]
#index the individuals
```

```
rownames(dp1_sample) <- rownames(gt_sample1)
#index the markers (SNPs)

dp2_sample <- do.call(cbind,lapply(2:col_n_sample,function(j){
  as.numeric(gsub('(.+)(:)(.+)(:)(.+)(:)(.+)(:)(.+)',
                  '\\5',gt_sample2[,j]))
}))
colnames(dp2_sample) <- colnames(gt_sample2)[2:col_n_sample]
#index the individuals
rownames(dp2_sample) <- rownames(gt_sample2)
#index the markers (SNPs)
#sizemat
#A matrix of total number of read counts
#the columns index the individual
#the rows index the markers(SNPs)

sizemat_sample <- rbind(dp1_sample,dp2_sample)
sizemat_sample #sizemat
#              JiNong99 N401 N402 N404 N405
#Arahy.01_90561        5   11   12   10   19
#Arahy.01_94948       34   28   22   22   11
#Arahy.01_91340        7    9   11   14   14
#Arahy.01_93550       12   14   23   17   24
#Arahy.01_94281       22   15   27   15   22
#Arahy.01_94568       17   24   17   17   17

class(sizemat_sample)
#check the data type
#"matrix" "array"

#extract information: refrence read counts "AD"
ad1_sample <- do.call(cbind,lapply(2:col_n_sample,function(j){
  as.numeric(gsub('(.+)(,)(.+)',
                  '\\1',
                   gsub('(.+)(:)(.+)(:)(.+)(:)(.+)(:)(.+)(:)(.+)(:)(.+)',
                        '\\3',
                       gt_sample1[,j])))
}))

colnames(ad1_sample) <- colnames(gt_sample1)[2:col_n_sample]
#index the individuals
rownames(ad1_sample) <- rownames(gt_sample1)
#index the markers (SNPs)
```

```
ad2_sample <- do.call(cbind,lapply(2:col_n_sample,function(j){
  as.numeric(gsub('(.+)(,)(.+)',
                  '\\1',
                  gsub('(.+)(:)(.+)(:)(.+)(:)(.+)(:)(.+)',
                       '\\3',
                       gt_sample2[,j])))
}))
colnames(ad2_sample) <- colnames(gt_sample2)[2:col_n_sample]
#index the individuals
rownames(ad2_sample) <- rownames(gt_sample2)
#index the markers (SNPs)

#refmat
#A matrix of reference read counts
#the columns index the individual
#the rows index the markers(SNPs)

refmat_sample <- rbind(ad1_sample, ad2_sample) #refmat

ploidy <- 4 # ploidy of peanut

genotyping_peanut_sample <- multidog(refmat = refmat_sample,
#utilise multidog to do genotyping
                                     sizemat = sizemat_sample,
                                     ploidy = ploidy,
                                     nc = NA)
```

```
   |                          *.#,%
  |||                        *******/
 |||||||   (**..#**.             */   **/
|||||||||   */*****************************/*%
  |||   &****..,*.************************/
  |||    (....,,,*,...****%********/(******
  |||            ,,****%////,,,,./.****/
  |||              /**//        .*///....
  |||              .*/*/%#        .,/   .,
  |||            , **/   #%        .*    ..
  |||                        ,,,*
```

```
Working on it...Loading required package: foreach
Loading required package: future
Loading required package: rngtools
done!>
```

2. Genotyping quality filter script

```
####Quality filter.....
load("/Users/han/Downloads/updog_res.RData")
nrow(res) #res - updog result ##935231

#missing data <= 0.01
#filter missing data <= 0.01
missing_data <- rowMeans(is.na(res)) #NA ratio per line
res1 <- data.frame(cbind(res,missing_data)) #res1 - updog result with NA ratio per line
nrow(res1) ## 935231
res2 <- res1[res1$missing_data <= 0.1, ] #filter res based on missing data <= 0.1
nrow(res2) ## 935231

#MAF >= 0.05
#filter res by MAF >= 0.05
#first calculate MAF
AA <- apply(res, 1, function(x) sum(x == "0", na.rm = TRUE))
sum(is.na(AA)) #0
Aa <- apply(res, 1, function(x) sum(x == "1", na.rm = TRUE))
aa <- apply(res, 1, function(x) sum(x == "2", na.rm = TRUE))

total_alleles <- (AA+Aa+aa)*2
A_freq <- round((AA*2+Aa)/total_alleles, 3)
a_freq <- round((aa*2+Aa)/total_alleles, 3)
res1 <- data.frame(cbind(AA, Aa, aa, total_alleles, A_freq, a_freq))
res1$MAF <- pmin(res1$A_freq,res1$a_freq)
#Choose the smaller value between A_freq and a_freq)
res2 <- res1[res1$MAF >= 0.05, ]
#res2 - filtered res1 based on MAF >= 0.05
nrow(res2) # 578088
res2[1:10,]
snps <- rownames(res2)
f_res <- res[rownames(res) %in% snps,]
nrow(f_res)
f_res[1:10,1:3]
save(snps, file = "snps.RData")
save(f_res, file = "f_res.RData")

#linux--beagle result
library(vcfR)
beagle <- read.vcfR("out.s353.gt.vcf.gz")
```

```
chrom <- getCHROM(beagle)
pos <- getPOS(beagle)
ref <- getREF(beagle)
alt <- getALT(beagle)
id <- getID(beagle)
gt <- extract.gt(beagle)
beagle <- data.frame(cbind(chrom, id, pos, ref, alt, gt))
beagle$id <- paste(beagle$chrom, beagle$pos, sep = "_")
f_beagle <- beagle[beagle$id %in% snps,]
index_order <- match(snps, rownames(f_beagle))
f_beagle <- f_beagle[index_order, ]
save(f_beagle, file = "f_beagle.RData")
```

3. GHap Haplotyping scripts

```
#### TO get .samples file.....
##resource
##"f_res.RData"    filtered res > f_res
load("f_res.RData")
samples = data.frame(matrix(nrow = ncol(f_res), ncol = 0))
samples$population <- seq_len(nrow(samples))
samples$ID <- rownames(f_res)
#save as .samples
write.table(samples, file = "res.samples", sep = " ", row.names = FALSE, col.names = FALSE)

#### TO get .markers file.....
##resource:
##"f_beagle.RData" filrered beagle > f_beagle
load("f_beagle.RData")
markers <- f_beagle[,1:5]
markers <- as.data.frame(markers)
markers$chrom <- as.numeric(gsub("Arahy\\.(\\d+)", "\\1", markers$chrom))
markers$pos <- as.numeric(markers$pos)
markers <- markers[order(markers$chrom, markers$pos), ]

#save as .markers
write.table(markers, file = "res.markers", sep = " ", row.names = FALSE, col.names = FALSE)

#### TO get .phase file.....
##resource
##"f_res.RData"    filtered res > f_res
##"f_beagle.RData" filrered beagle > f_beagle

#load data
load("f_res.RData")    #"f_res"
load("f_beagle.RData") #"f_beagle"

##check data.....
#count snps
nrow(f_res)          #578088
ncol(f_res)          #353 (individuals)
nrow(f_beagle)       #578088
ncol(f_beagle)       #358 (include 5 extra column)
f_beagle <- f_beagle[, -(1:5)]

#check missing data and position
sum(is.na(f_res)) #306
```

```
sum(is.na(f_beagle))
mising_data_position <- which(is.na(f_res))
mising_data_position

##check reversal and do adjustment.....
#make a new data.frame shows the 0 and 0|0 percentage per row in f_res and f_beagle
#calculate the 0 percentage per snp in f_res
#calcilate the 0|0 percentage per snp in f_beagle
#Calculate the 0 percent difference for each snp
#set a threshold: 0.1
#if the difference larger than 0.1,
#then i consider the alt and ref is reversal in f_res and do adjustment
row_names <- rownames(f_res)
zero_percentage <- data.frame(matrix(nrow = length(row_names), ncol = 0))
rownames(zero_percentage) <- row_names

f_res <- as.data.frame(f_res)
f_beagle <- as.data.frame(f_beagle)

zero_percentage$f_res <- apply(f_res, 1, function(row) sum(row == 0)/353)
zero_percentage$f_beagle <- apply(f_beagle, 1, function(row) sum(grepl("0\\|0", row))/353)
zero_percentage$abs_diff <- abs(zero_percentage$f_res - zero_percentage$f_beagle)
nrow(zero_percentage) #578088

f_zero_percentage <- subset(zero_percentage, abs_diff > 0.1)
nrow(f_zero_percentage) #27
matching_rows <- rownames(f_zero_percentage)
for (row_name in matching_rows) {
  row_index <- which(rownames(f_res) == row_name)
  f_res[row_index, f_res[row_index, ] == 0] <- 2
  f_res[row_index, f_res[row_index, ] == 2] <- 0
}

##adjust f_res rest content.....
#change 0 to 0|0
#change 2 to 1|1
#Replace <NA> in f_res with the content in the corresponding position of f beagle
#Replace 1 with the content in the corresponding position of f beagle 1|0 or 0|1
f_res[f_res == 0] <- "0|0"
f_res[f_res == 2] <- "1|1"
for(i in seq_along(f_res)) {
  . <- is.na(f_res[[i]])
  f_res[[i]][.] <- f_beagle[[i]][.]
}
```

```
sum(is.na(f_res)) #ckeck NAs in f_res #0

for (i in seq_along(f_res)) {
  . <- f_res[[i]] == 1
  f_res[[i]][.] <- f_beagle[[i]][.]
}

##create matrix m*2n, two column per individual.....
#separate the columns in f_res
library(tidyr)
phase <- f_res
for(i in colnames(phase)) {
  phase <- separate(phase, col = i, into = c(paste0(i, "_1"), paste0(i, "_2")))
}

#order
phase <- phase[row.names(markers),]

##save as .phase file
write.table(phase, file = "res.phase", sep = " ", row.names = FALSE, col.names = FALSE)

#### run GHap.....
library(GHap)
# Compress phase data using file names
ghap.compress(samples.file = "res.samples",
              markers.file = "res.markers",
              phase.file = "res.phase",
              out.file = "res")

# Load data using file names
phase <- ghap.loadphase(samples.file = "res.samples",
                        markers.file = "res.markers",
                        phaseb.file = "res.phaseb")

# Generate blocks of 5 markers sliding 5 markers at a time
blocks.mkr <- ghap.blockgen(phase, windowsize = 5, slide = 5, unit = "marker")

# Generate matrix of haplotype genotypes
ghap.haplotyping(object = phase, blocks = blocks.mkr,
                 outfile = "res_pop", binary = TRUE)

ghap.haplotyping(object = phase, blocks = blocks.mkr,
                 outfile = "res_pop_nb", binary = FALS
```

4. GHap GWAS script

```
#####1. Perform res_pop, which have 5snps per haplo blocks>>>>>>>>>>>>>>>>>>>>>>>>
##Manipulating haplo objects
# Load haplotype genotypes using prefix
haplo <- ghap.loadhaplo("res_pop")
# Convert to plink
ghap.hap2plink(haplo, outfile = "res_pop")
## Association analysis
#loads plink data
plink <- ghap.loadplink("res_pop")
#load phenotype data
df <- read.table(file = "res.phenotype.txt", header = T)
#cpmpute genomix relationship matrix
#introduce sparsity to hlpe with matrix inversion
K <- ghap.kinship(plink, sparsity = 0.01, type = 1)
K1 <- ghap.kinship(plink, sparsity = NULL, type = 1)
#perform GWAS
#Method 1.....
#without kinship and rep
gwas_pop_1 <- ghap.assoc(object = plink,
                         formula = pheno ~ 1 +(1|id),
                         data = df,
                         covmat = NULL,
                         ngamma = 100, nlambda = 100, recalibrate = 0.01)
ghap.manhattan(data = gwas_pop_1, chr = "CHR", bp = "BP", y = "LOGP")
#method 2.....
#without rep
gwas_pop_2 <- ghap.assoc(object = plink,
                         formula = pheno ~ 1 +(1|id),
                         data = df,
                         covmat = list(id = K),
                         ngamma = 100, nlambda = 100, recalibrate = 0.01)
ghap.manhattan(data = gwas_pop_2, chr = "CHR", bp = "BP", y = "LOGP")
#method 3.....
# Perform GWAS on repeated measures
# Use grammar-gama approximation
# Recalibrate top 1 percent variants
df$rep <- df$id
gwas_pop_3 <- ghap.assoc(object = plink,
                  formula = pheno ~ 1 + (1|id) + (1|rep),
                  data = df,
                  covmat = list(id = K, rep = NULL),
                  ngamma = 100, nlambda = 1000, recalibrate = 0.01)
```

```
ghap.manhattan(data = gwas_pop_3, chr = "CHR", bp = "BP", y = "LOGP")
```