Journal of Advanced Research 42 (2022) 237–248

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Original Article

Genomic insights into the genetic signatures of selection and seed trait loci in cultivated peanut

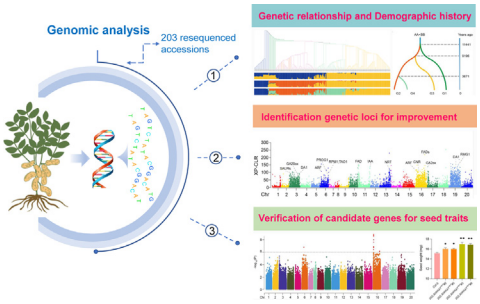
Yiyang Liu [a,1], Libin Shao [b,1], Jing Zhou [b,1], Rongchong Li [a], Manish K. Pandey [c], Yan Han [d], Feng Cui [a], Jialei Zhang [a], Feng Guo [a], Jing Chen [e], Shihua Shan [e], Guangyi Fan [b,f], He Zhang [f], Inge Seim [g,h], Xin Liu [f], Xinguo Li [a,*], Rajeev K. Varshney [c,i,j,*], Guowei Li [a,d,*], Shubo Wan [a,*]

[a] Provincial Key Laboratory of Crop Genetic Improvement, Ecology and Physiology, Institute of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Ji'nan 250100, Shandong Province, China
[b] BGI-Qingdao, BGI-Shenzhen, Qingdao, Shandong Province, China
[c] Center of Excellence in Genomics and Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad 502324, India
[d] College of Life Sciences, Shandong Normal University, Ji'nan 250014, Shandong Province, China
[e] Shandong Peanut Research Institute, Qingdao 266000, China
[f] State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China
[g] Integrative Biology Laboratory, College of Life Sciences, Nanjing Normal University, Wenyuan Road, Nanjing 210023, China
[h] School of Biology and Environmental Science, Queensland University of Technology, Brisbane 4000, Australia
[i] The UWA Institute of Agriculture, the University of Western Australia, Perth, WA 6001, Australia
[j] State Agricultural Biotechnology Centre, Centre for Crop and Food Innovation, Murdoch University, Murdoch, Western Australia, Australia

HIGHLIGHTS

- The study presents a large-scale SNP data set from the whole-genome resequencing of 203 cultivated peanut accessions.
- Population structure and demographic history are investigated.
- Signatures of selection occurred during peanut improvement breeding are demonstrated.
- Candidate genes associated with seed traits are identified by GWAS and transgenic experiments.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 2 November 2021
Revised 25 January 2022
Accepted 28 January 2022
Available online 1 February 2022

Keywords:
Peanut
Genomic diversity
Evolution
GWAS
Seed traits

ABSTRACT

Introduction: Cultivated peanut (*Arachis hypogaea* L.) is an important oil crop for human nutrition and is cultivated in >100 countries. However, the present knowledge of its genomic diversity, evolution, and loci related to the seed traits is limited.
Objectives: Our study intended to (1) uncover the population structure and the demographic history of peanuts, (2) identify signatures of selection that occurred during peanut improvement breeding, and (3) detect and verify the functions of candidate genes associated with seed traits.
Methods: We explored the population relationship and the evolution of peanuts using a largescale single nucleotide polymorphism dataset generated from the genome-wide resequencing of 203 cultivated peanuts. Genetic diversity and genomic scan analyses were applied to identify selective loci for genomic-selection breeding. Genome-wide association studies, transgenic experiments, and RNA-seq were employed to identify the candidate genes associated with seed traits.

Peer review under responsibility of Cairo University.
* Corresponding authors.
E-mail addresses: xinguol@163.com (X. Li), R.K.Varshney@cgiar.org (R.K. Varshney), liguowei@sdnu.edu.cn (G. Li), wanshubo2016@163.com (S. Wan).
[1] Authors contributed equally.

https://doi.org/10.1016/j.jare.2022.01.016

Results: Our study revealed that the 203 resequenced accessions were divided into four genetic groups, consistent with their botanical classification. Moreover, the var. *peruviana* and var. *fastigiata* subpopulations have diverged to a greater extent than the others, and var. *peruviana* may be the earliest variant in the evolution from tetraploid ancestors. A recent dramatic expansion in the effective population size of the cultivated peanuts *ca.* 300–500 years ago was also noted. Selective sweeps underlying quantitative trait loci and genes of seed size, plant architecture, and disease resistance coincide with the major goals of improved peanut breeding compared with the landrace and cultivar populations. Genome-wide association testing with functional analysis led to the identification of two genes involved in seed weight and seed length regulation.

Conclusion: Our study provides valuable information for understanding the genomic diversity and the evolution of peanuts and serves as a genomic basis for improving peanut cultivars.

Introduction

The cultivated peanut or groundnut (*Arachis hypogaea* L.) is an allotetraploid (AABB, $2n = 4 \times = 40$) cultivated worldwide from the tropical to temperate zones, providing approximately 46 million tons annually [1]. The genus *Arachis* L. includes 81 described species that are further divided into 31 sections, and *Arachis hypogaea* L. belongs to the *Arachis* section [2]. *A. hypogaea* L. is classified into two subspecies and six varieties based on morphological features: var. *fastigiata*, var. *vulgaris*, var. *peruviana,* and var. *aequatoriana*, belong to *ssp. fastigiata*, whereas var. *hypogaea* and var. *hirsuta* belong to *ssp. hypogaea* [3]. The primary and secondary centers of origin and diversity of *A. hypogaea* have been proposed in South America [4]. The genetic variation among peanut tetraploids is expected to decrease owing to polyploidization. A domestication bottleneck has long been thought to complicate plant breeding [5], and it is estimated that peanuts experienced 6 such bottlenecks during their evolution and domestication [6]. However, our knowledge of its genetic diversity and evolution remains limited. Specifically, our understanding of selective sweep signatures during peanut domestication and breeding needs to be improved.

The development of peanut cultivars has dramatically enhanced the yield, quality, and adaptation of the crop to diverse growth conditions. The average peanut yield worldwide has increased by approximately 80% during the last 60 years (http://www.fao.org). However, intensive breeding has led to the loss of diversity. Understanding the genomic basis of modern breeding may provide key insights for further improvement and adaptation by comparing the genetic diversity between landraces and modern cultivars. Previous studies using a 58K single nucleotide polymorphism (SNP) array revealed low genetic diversity in cultivated and landrace peanuts [7,8]. Nonetheless, considering the low number of SNPs captured and the large tetraploid genome of the peanut, genome-wide studies are warranted to comprehensively explore the genetic diversity of peanut cultivars and landraces. Several genes or quantitative trait loci (QTLs) related to yield have been identified, which further boosted this effort [9]. However, seed size-associated genes have not been adequately identified and validated experimentally.

In this study, we performed whole-genome resequencing of 203 accessions from across the world to assess the population structure and demographic history of peanuts, identify the signatures of selection that occurred during peanut breeding, and detect the function of candidate genes associated with seed traits. This endeavor enabled the identification of multiple candidate loci and genes for agronomic traits and provided genetic resources of cultivated peanuts that are likely to facilitate peanut cultivar breeding in the future.

Materials and methods

Plant materials a total of 203 individual peanut cultivar accessions were collected from across the world, including five varieties from 31 countries and 25 wild *Arachis* accessions provided by ICRISAT (Table S1). A single-seeding precision sowing method [10,11], with 237,000 seedlings per hectare, was used as the study site for plant cultivation. An area of three mulches with a 2-m area was used for each accession. Three biological replicates were used to investigate the phenotype at different regions (May-September 2018 at Jinan, Shandong, 36°40′N, 117°00′E, April-August 2019 at Changsha, Hunan, 28°12′N, 112°59′E and November 2019 to 2020 February at Ledong, Hainan 18°45′ N, 109°10′E). DNA was extracted from the leaves of each accession. The tissues for transcriptomic analysis include leaves, roots, stems, flowers, and seeds at 4 different stages [seed 1, 30 days after flowering (DAF); seed 2, 40 DAF; seed 3, 50 DAF; seed 4, 60 DAF] were collected and all the samples were transferred into liquid nitrogen immediately and stored at −80 °C until DNA or RNA extraction.

Whole-genome resequencing and variant identification

Genomic DNA was extracted from young leaves and subjected to library construction and amplification as per the standard protocols specified by the Illumina HiSeq 4000 Sequencer and BGISEQ-500. To explore the genetic variation in cultivated peanut, we re-sequenced 203 accessions representing five botanical varieties from 31 countries and 25 wild *Arachis* accessions. The genome of cv. Tifrunner was used as a reference [12]. We used SOAPnuke (v1.6.0) to remove the raw reads containing 1) sequencing adapter; 2) low-quality base ratio (base quality $\leq$ 12) of > 50%; 3) N (unknown base) ratio of > 10%. Data analysis and variation calling was further performed based on the clean data [13].

After data filtration, we employed the SentieonDNAseq software to detect SNPs (https://www. sentieon.com/products) [14], which is a high speed of reads mapping and SNPs calling toolkit for MPS reads. First, the clean reads of all individuals were mapped to the genome sequences of the cultivated peanut (https://www.peanutbase.org/data/v2/Arachis/hypogaea/genomes/Tifrunner.gnm1.KYV3). We used the Sentieon BWA model for alignment and then called SNPs using the Sentieon Haplotyper model (the same algorithm with GATK). Finally, we used GATK to integrate multiple individual SNP sets of gVCF files to a final population SNP set with the VCF format [15]. For the quality control of SNPs, we used a SelectVariants model with the following parameters in GATK to remove low quality SNPs: QualByDepth (QD $\geq$ 2), FisherStrand (FS $\leq$ 60), RMSMappingQuality (MQ $\geq$ 40), MappingQualityRankSumTest (MQRankSum $\geq$ -12.5), ReadPos-

RankSum (ReadPosRankSum ≥ -8.0) and StrandOddsRatio (SOR > 3.0).

Phylogenetic analysis and population structure

To perform the population structure analysis, we kept biallelic (only two alleles) SNPs and screened high-quality SNPs using PLINK (ver. 1.90) [16]. SNPs and individuals that met any of the following stringent quality control parameters were removed: if 1) minor allele frequency (--maf) < 0.05; 2) individuals call rate (--mind) ≤ 0.2; 3) missing genotype frequency for SNP (--geno) ≥ 0.05; 4) Hardy-Weinberg equilibrium *P*-value (--hwe) < 1e-6. The remaining markers were used for further population analysis. We performed principal component analysis (PCA) with the whole high-quality SNPs (6,686,684 SNPs) using the smartPCA program from the EIGENSOFT package [17], while the top 3 eigenvectors were plotted in two dimensions. In addition, a neighbor-joining phylogenetic tree was constructed using SNPs at fourfold degenerated sites (25,572 SNPs) by the MEGA7 with 1000 bootstraps. The tree was displayed using the EvoIView (v3) [18]. The population structure was constructed using the ADMIXTURE software [19], which is based on a variational Bayesian framework for posterior inference. To identify the best genetic cluster *K*, we tested the number of *K* values from 2 to 5 with 1000 iterations for each run.

Population genetic analysis

Fixation statistics (F_{ST}) and nucleotide diversity (π) in 50-kb non-overlapping sliding window were calculated using the program VCFtools (v0.1.13) along each chromosome for each group according to previous analysis [5,20].

Linkage disequilibrium

To explore the linkage disequilibrium (LD), any of the two SNP's max distance (kb) was set to 300 in order to calculate the correlation coefficient (r^2) using the PopLDdecay 3.40 [21]. LD decay statistics were calculated for different groups, while the LD decay graphs were plotted with the parameter-MaxDist 300.

Demographic history and divergence time

We applied SMC++ v. 1.15.4 to estimate the demographic history that seemed powerful for recovering the effective population size history with short timescales based on whole-genome sequence data [22]. Four groups of 203 individuals were calculated as the effective population size separately. A generation time per year and a rate of 1.68×10^{-8} mutations/nucleotide/year [23,24] were used to remake the scaled times and effective population sizes into real times and sizes, while other parameters were set as default with outlier individuals excluded.

We used fastsimcoal [25] to infer divergence time in peanuts. First, four-fold Degenerate Synonymous Site (4DTv) were filtered by vcftools (version 0.1.17). Then, the easySFS.py script was used for converting vcf to construct the site frequency spectrum. Finally, fastSimcoal2 (version 2.6.0.3–14.10.17) with "-t fsc.tpl -n 10,000 -m -e fsc.est -M -L 20 -c12 -B 12 -q" was used as the parameter to estimate demographic parameters from the site frequency spectrum.

Identification of selective signatures

A cross-population composite likelihood ratio test (XP-CLR) [26] and population fixation statistics (F_{ST}) were employed to detect selective signatures between the improved cultivars and other landraces. The XP-CLR value was tested with the parameter (-w1 0.0005 100 100 1 -p0 0.7) for each chromosome by the mean likelihood score in 500 kb sliding windows with a step size of 50 kb across the genome. F_{ST} was calculated using the VCFtools (v0.1.13) in a 500-kb sliding window with a 50-kb step [27]. The average F_{ST} of all sliding windows was regarded as the value at the whole-genome level across different groups. Sliding windows with the top 5% highest values both in F_{ST} and XP-CLR tests were regarded as candidate regions with strong selection and annotated genes residing in these regions were considered candidate selected genes. GO term enrichment analyses were performed for candidate selected genes using *perl* module GO.

Phenotyping for agronomic traits

For phenotyping, the 203 accessions were planted in three environmental conditions in 3 years (2017 to 2019) at 3 different provinces (Shandong, Hunan, and Hainan). The phenotypic value of seed traits was the mean of 10 measurements. Seed weight was obtained by weighing 20 seeds for each accession. Considering the traits related to seeds (weight, length, and width) are sensitive to the environment, we used the data from lines that are normally developed and matured in these three locations for the downstream genome-wide association studies (GWAS).

GWAS analysis and candidate genes identification

The best linear unbiased prediction (BLUP) was produced for genetic evaluation to integrate multiple environmental data, remove environmental deviations, and obtain real genetic phenotypes of the individuals [28,29]. To accurately describe the individual traits, we used BLUP to recalculate the seed traits values by using an R script, with the 'lme4' package of the R v3.4 software (www.r-project.org) to calculate BLUPs values, with location and year serving as random effects in the model lmer (phenotypes ~(1|loc) +(1|year) +(1|lines) +(1|year: lines). The BLUP values were used as individual phenotypes for the GWAS analysis. In total, 1,291,801 SNPs with a minor allele frequency (MAF) of ≥ 0.05 and a missing data rate of ≤ 5% in the entire population were used for GWAS analyses. To avoid spurious associations, Plink software was used to generate a PCA matrix. The Efficient Mixed-Model Association eXpedited (EMMAX) was used to test trait-SNP associations, which includes the kinship matrix as a random effect in the mixed effect model, and EMMAX was represented as: Y = SNP + PCs + Kinship + e. We used 10 PCs as the population structure, and Kinship matrix between the individuals represents the relationship. SNPs and PCs are set as fixed effects, while kinship is a random effect. The analysis process is as follows: (1) PCA is required as fixed effect (-c): plink --bfile [bed_prefix] --maf 0.05 --geno 0.05 --chr-set 20 --allow-extra-chr --pca 10 --out [tped_prefix]. (2) Use constant as a random effect (-k): emmax -v -d 10 -t [tped_prefix] -p [pheno_file] -k [kin_file] -o [out_prefix]. (3) GWAS analysis: emmax -d 10 -t [tped_prefix] -p [pheno_file] -c [PCA:10] -k [hBN.kinf] -o [out_prefix] [30]. The $-\log_{10}P > 6$ was selected as the significance threshold of the associated SNPs. To further identify reliable significant signals in the GWAS results, only the LD blocks containing at least one significant and one suggested SNPs were considered as the significant loci. Each of the candidate genes associated with these significant loci was extracted for functional annotation, and the homologs from closely related species or the model species *Arabidopsis thaliana* or rice were further identified by BLASTP. Based on the abovementioned results, the candidate genes associated with each trait were further analyzed.

Transcriptome sequencing and analysis

The seeds at 50 days after flowering (seed 3) were collected from randomly selected 5 large-seed size and 5 small-seed size accessions. Total RNA was extracted and purified using the QIAGEN RNeasy Plant Mini Kit (Qiagen, USA) according to the manufacturer's protocols. The quality and concentration of RNA were detected by the NanoDrop One UV–Vis Spectrophotometer (Thermo Fisher Scientific, USA). The paired-end sequencing with 150 bp was conducted on the BGISEQ-500 Sequencer (BGI, China). To generate the clean data, the raw data were further removed reads containing poly-N, adapters, and low-quality. The clean data were then mapped to the genome sequences of the cultivated peanut (https://www.peanutbase.org/data/v2/Arachis/hypogaea/genomes/Tifrunner.gnm1.KYV3). For gene expression quantification in different tissues, fragments per kilobase of exon model per million reads mapped (FPKM) values were calculated using the HISAT2 (v2.1.0) and Cufflinks (v2.2.1).

qRT-PCR for gene expression analysis

The total RNA tissues from cv. Tifrunner included roots, stems, leaves, flowers, and seeds at seed 1–4 stages. All RNA samples were extracted using the TIANGEN RNAprep pure Plant Kit (Tiangen, China) according to the manufacturer's protocol and reverse transcribed using the PrimeScript RT Reagent Kit with gDNA Eraser (TaKaRa) [31]. qRT-PCR was performed with the SYBR Premix Dimer Eraser (TaKaRa). The relative gene expression was analyzed using the $2^{-\Delta\Delta CT}$ method for samples from at least three replicates. The qRT-PCR Primers were designed by Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast) and sequences are listed in Table S10.

Gene cloning and plant transformation

Gene cloning and plant transformation was performed following the previous methods [31]. The protein coding sequences of the related genes were obtained through PCR using cDNA derived from seeds of cv. Tifrunner and the corresponding varieties with SNP alleles. The amplified products were further cloned into the PHB vector driven by the cauliflower mosaic virus (CaMV) 35S promoter. The resulting constructs were transformed into *A. thaliana* (Col-0) by *Agrobacterium tumefaciens* GV3101 and selected with Basta. Homozygous lines of transgenic plants were used in this study. Five hundred seeds from WT and homozygous transgenic Arabidopsis lines were randomly counted and weighed using an electronic balance (Mettler-Toledo, Switzerland). For size measurement, the seeds were photographed and measured using ImageJ software. The primers sequences used for gene cloning are listed in Table S10.

Total lipid content and fatty acid composition analysis

We determined the total lipid content in the seeds as suggested by a previous report [32]. Briefly, 50 mg of dry seeds were vortexed and saponified by refluxing with methanol (50%) containing 5% sodium hydroxide for 1 h. Next, 2 mL chloroform was added to the tube and vortexed for 10 min. After centrifugation of the mixture at 5000 g for 5 min at room temperature, the lower phase was transferred to a dry weighed glass tube; 1 mL of hexane was added to the remaining upper layer, and the tube was vortexed for 10 min. After centrifugation, the upper phase was collected and added to the transferred chloroform solution. Finally, the mixture was dried under nitrogen flow, and the total lipid content was calculated. Three biological replicates were performed for each line.

Fatty acid composition analysis was performed as previously described with the Agilent 6890 N [33] using 0.3 g of mature dry seeds. The extracted FAMEs were quantified by gas chromatography-mass spectrometry (GC–MS) on the HP-88 capillary column (30 × 0.25 × 0.25 mm) with the DAOJING GC-2010. Triheptadecanoin was used as the triacylglycerol internal standard, and fatty acid compositions were quantitatively analyzed using the internal standard.

Results

Sequencing and genomic variations

Domesticated peanut cultivars are widely grown and showed high geographical diversity, albeit their genomic variation and the genetic basis of their agriculturally important traits remain largely unexplored. To examine the population structure and genomic variation of cultivated peanuts, we collected and sequenced 203 accessions including 153 landraces and 50 improved cultivars, from 31 countries. These accessions represented five botanical varieties [var. *fastigiata* (80), var. *hypogaea* (56), var. *vulgaris* (47), var. *peruviana* (11), and var. *hirsuta* (9)] (Fig. 1A, Table S1). We obtained 7.19 Tb of high-quality reads, with an average of 14.16-fold depth for each accession. The 25 wild *Arachis* accessions with 30.9 Gb average resequencing data for each accession as a control group. By using the chromosome-level assembled genome of cv. Tifrunner as the reference genome, we identified 6,686,684 high-quality SNPs from the 203 cultivated and 25 wild accessions. Among the 5,439,320 SNPs in cultivated accessions, 210,240 were located in the upstream or downstream genic regions, and 3,369,870 were located in intergenic regions. The protein-coding regions harbored 321,406 SNPs, including 47,995 nonsynonymous, 1,125 splicing, 155 stop-loss, and 2,397 stop-gain SNPs that caused amino acid changes, elongated transcripts, and premature stopping, respectively (Table S2). Moreover, the number of SNPs in the B subgenome (3,224,418) was approximately 1.5-fold higher than that in the A subgenome (2,172,110), which is consistent with the ratio of the B/A subgenome.

Population structure

We performed multiple analyses to explore the genetic components and relationships of the sequenced peanut populations. We first constructed a neighbor-joining phylogenetic tree with the wild *Arachis* species as an outgroup using SNPs at four-fold degenerate sites. The phylogenetic tree revealed that the var. *peruviana* accessions initially split from the wild *Arachis* species at the root, followed by the var. *hypogaea* plus var. *hirsuta* accessions. Subsequently, the var. *vulgaris* accessions and var. *fastigiata* accessions were further divided into two genetic groups (Fig. 1B), which supports the current cultivated peanut classification based on distinctive botanical features. Some accessions with sporadic distribution could not be grouped by their designation. This observation may reflect the misclassification during collection or the inconsistency of genetics and phenotypes. We further performed model-based clustering analysis using ADMIXTURE. The results indicated that the same groupings were identified among these accessions, which is consistent with the phylogenetic analysis (K = 4), in which the accessions were categorized into four clusters, namely, var. *peruviana* (Group I, G1), var. *hypogaea* plus var. *hirsuta* (Group II, G2), var. *vulgaris* (Group III, G3) and var. *fastigiata* (Group IV, G4) (Fig. 1B). A principal component analysis (PCA) of the accessions revealed four clusters corresponding to the lineages of groups I-IV, which corroborates the phylogenetic and ADMIXTURE results

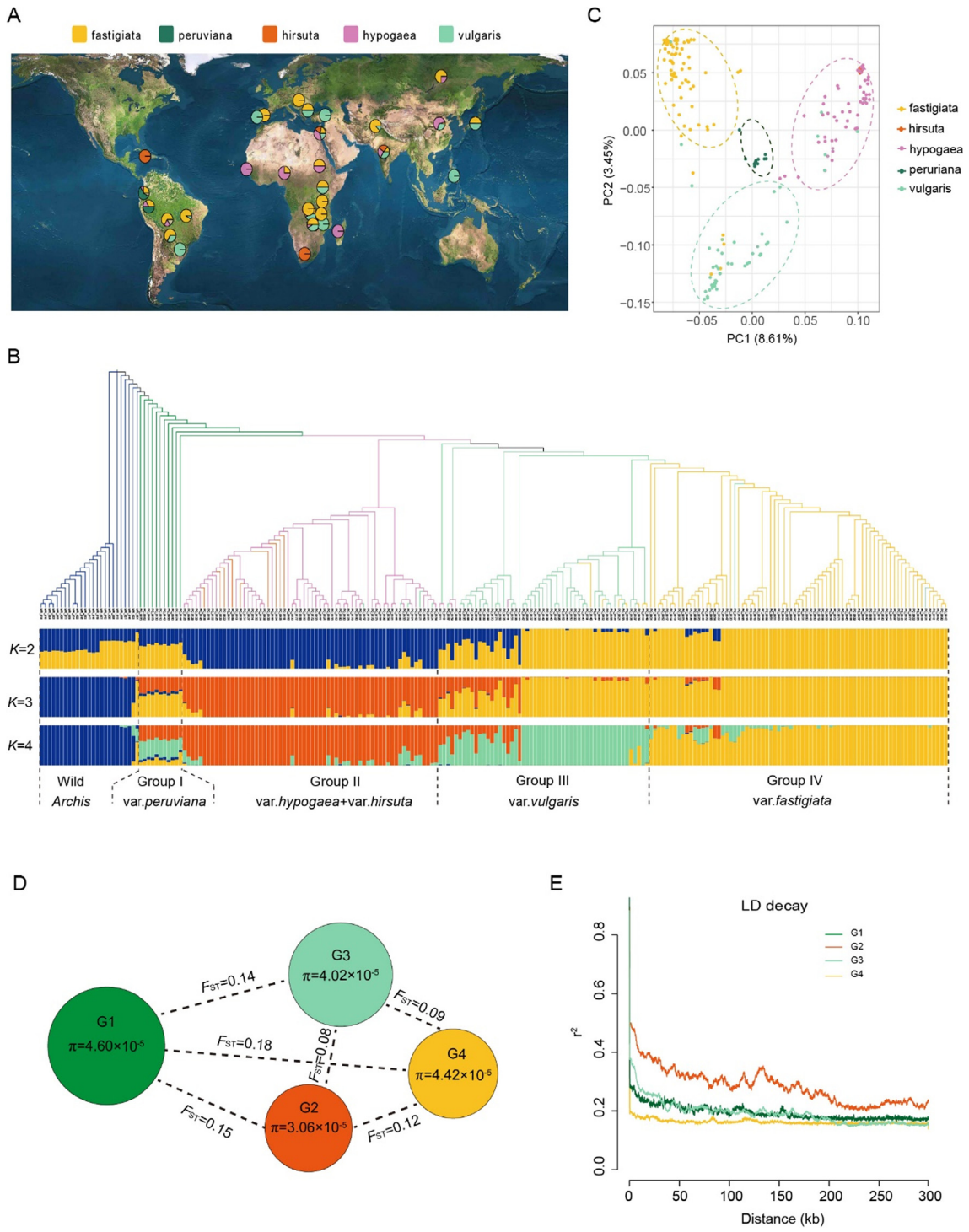


Fig. 1. Geographical distribution and population structure of 203 resequenced accessions. (A) The geographical distribution of all 203 accessions. The five varieties are indicated in different colors. The proportion of each pie chart indicates the ratio of the different variety types in a certain country. The colors varied with the variety. **(B)** Neighbor-joining phylogenetic tree of all 203 accessions with wild *Arachis* accessions as outgroup and population structure analysis with K = 2 to 4. The colors of branches in the phylogenetic tree represent different variety types (consistent with the colors shown in A). **(C)** Principal component analysis of all 203 accessions. The first two principal components (PC1 and PC2) are shown and dot colors represent the different accessions. The colors of the virtual circle match the structure grouping. **(D)** The genetic diversity (π) and population differentiation (F_{ST}) across the 4 groups. The values in the circles represent the genetic diversity of the groups (green, tangerine, cyan, and yellow circles represent G1 to G4 groups, respectively), and the F_{ST} values between the groups are shown. The radius of the pie represents the genetic diversity value and the length of dashed line represents the F_{ST} value between groups. **(E)** Linkage disequilibrium decay was estimated from different peanut groups. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(Fig. 1C). It is therefore apparent that most of the improved cultivars received genetic contributions from var. *peruviana*, var. *hypogaea*, and var. *vulgaris* species, whereas the landraces rarely received genetic contributions from other species. Interestingly, when K = 5, the improved cultivars clearly diverged from G2 and G3 (Fig. S1), which suggests that the improved varieties were derived from the crosses between different subspecies of var. *hypogaea* and var. *vulgaris*.

We calculated F_{ST} (fixation index values) to estimate the genetic distance between groups. The F_{ST} values between G1 (var. *peruviana*) and the other three groups were higher than those of the other groups. The highest diversity (F_{ST} = 0.18) was observed between G1 (var. *peruviana*) and G4 (var. *fastigiata*), and the lowest was observed between G2 (var. *hypogaea* plus var. *hirsuta*) and G3 (var. *vulgaris*) (F_{ST} = 0.08). Hence, it could be inferred that var. *peruviana* and var. *fastigiata* subpopulations have diverged to a greater extent than the others. The varieties var. *hypogaea* plus var. *hirsuta* and var. *vulgaris* demonstrated a relatively close genetic relationship. These results also signified that the F_{ST} values displayed a botanically correlated pattern, which is consistent with the findings of the phylogenic analysis.

Moreover, we investigated the genetic diversity by calculating the nucleotide diversity (π) in each group. First, we explored the nucleotide diversity of the accessions from the so-called diversity center (South America), the transfer station (Africa), and the primary production region (Asia) [4]. The π value (4.44×10^{-4}) of the accessions from South America was the highest, the π value (3.55×10^{-5}) of the accessions from Asia was the lowest, and the π value of the accessions from Africa was intermediate (4.14×10^{-5}), which is in accordance with the migration route from South America to Africa and then Asia. Next, we compared the nucleotide diversity of different groups. The result revealed that G2 ($\pi = 3.06 \times 10^{-5}$) and G3 ($\pi = 4.02 \times 10^{-5}$) had lower nucleotide diversity than G4 ($\pi = 4.42 \times 10^{-5}$) and G1 ($\pi = 4.60 \times 10^{-5}$) (Fig. 1D). These results are in line with the expected values because var. *hypogaea* and var. *vulgaris* are the most widely grown varieties and are used primarily for breeding improvement, which possibly decreases their genetic diversity [34,35]. The LD decay rates were the highest in G2, followed by G3 and G1, and the lowest in G4 (Fig. 1E). This result is correlated with the level of nucleotide diversity. The LD extended further in G1 and G3 than in the others, which suggests a possible bottleneck during their breeding history.

Demographic history and divergence time

Although recent studies have helped clarify the origin of the cultivated peanut, the history of diversification of different botanical varieties is not well understood. Since peanut is regarded to have originated from a polyploidization event <10,000 years ago, we investigated the demographic history of the crop that resulted in the current populations using SMC++ (Fig. 2A). The demographic history with SNP data revealed that all four subpopulations experienced a slight decline in the effective population size (N_e) from the highest point *ca.* 10,000 years ago to a nadir *ca.* 1000 years ago. The domesticated peanut experienced a dramatic expansion in N_e, which started 300–500 years ago. Notably, the initial period of N_e expansion of peanuts coincided with the European colonial period in South America (1492–1832), which supports the view that cultivated peanuts originated in South America and were domesticated there; they were later were spread worldwide by the European colonialists.

We used fastsimcoal to infer the divergence time of the peanut varieties. A four-fold Degenerate Synonymous Site (4DTv) was employed to estimate the demographic parameters from the site frequency spectrum (Fig. 2B). The optical model indicated that G1 first diverged 11,441 years ago. Then, G3 emerged 9,196 years ago and, finally, G2 and G4 separated 3,671 years ago. This result is consistent with the date of origin estimated by demographic history analysis.

Selection signals under breeding improvements of peanut

Like other crop species, the cultivated peanut has undergone continuous selection because of domestication and intensive breeding events to maximize its yield and quality. We estimated the genetic diversity of 153 landraces and 50 improved varieties that were separated by approximately 20 years in breeding history (Table S1). The average π value of the landraces was 4.20×10^{-5}, which is higher than that of the improved cultivars (3.56×10^{-5}). Particularly, the LD decay rate of the peanut cultivars was markedly lower than that of the landraces (Fig. S2). These results indicate that the modern improved cultivars have experienced more artificial selections than the landraces.

We compared the modern improved cultivars with the landraces using genetic differentiation (F_{ST}) and XP-CLR (cross-population composite likelihood ratio) tests to identify the selection signals and candidate genes involved in breeding improvements. Totally, 637 sweep regions encompassing 3,413 genes, were identified by both methods (Fig. 4A; Table S3 and S4). Moreover, the number of selected genes in the B subgenome (2,299) was approximately two-fold higher than that in the A subgenome (1,084), which implies an asymmetric subgenome selection during

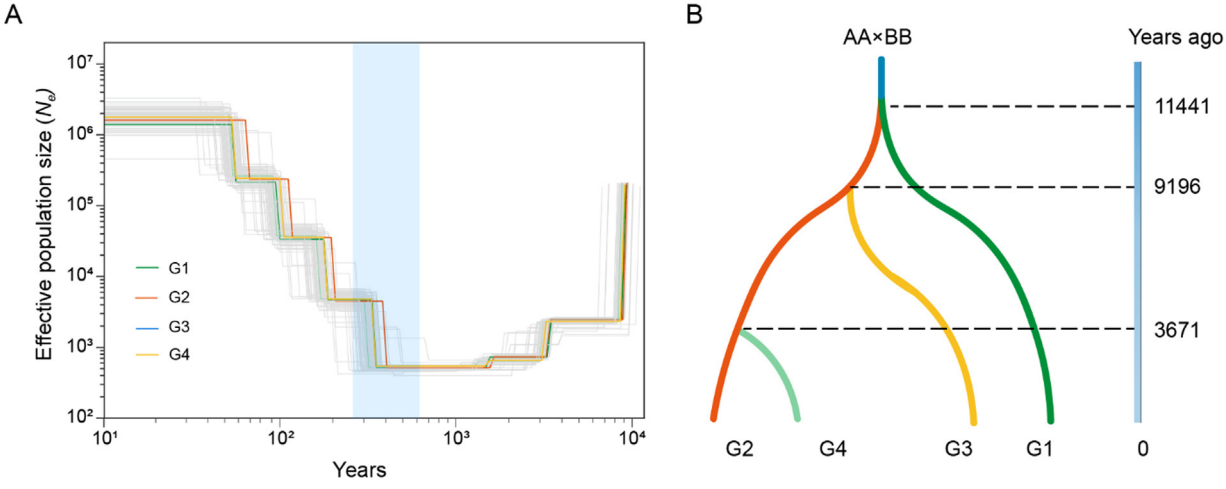


Fig. 2. Demographic history of peanut populations. (A) The demographic history was demonstrated by SMC++. The rebound in the effective population size 300–500 years ago is shaded in light blue color. The different colored solid lines indicate typically demographic history in each group. One generation per year and a mutation rate of 1.68×10^{-8} per generation were used to scale the real-time. **(B)** Schematic of divergence time using fastsimcoal2. The vertical axis indicates the estimated time of population divergence. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

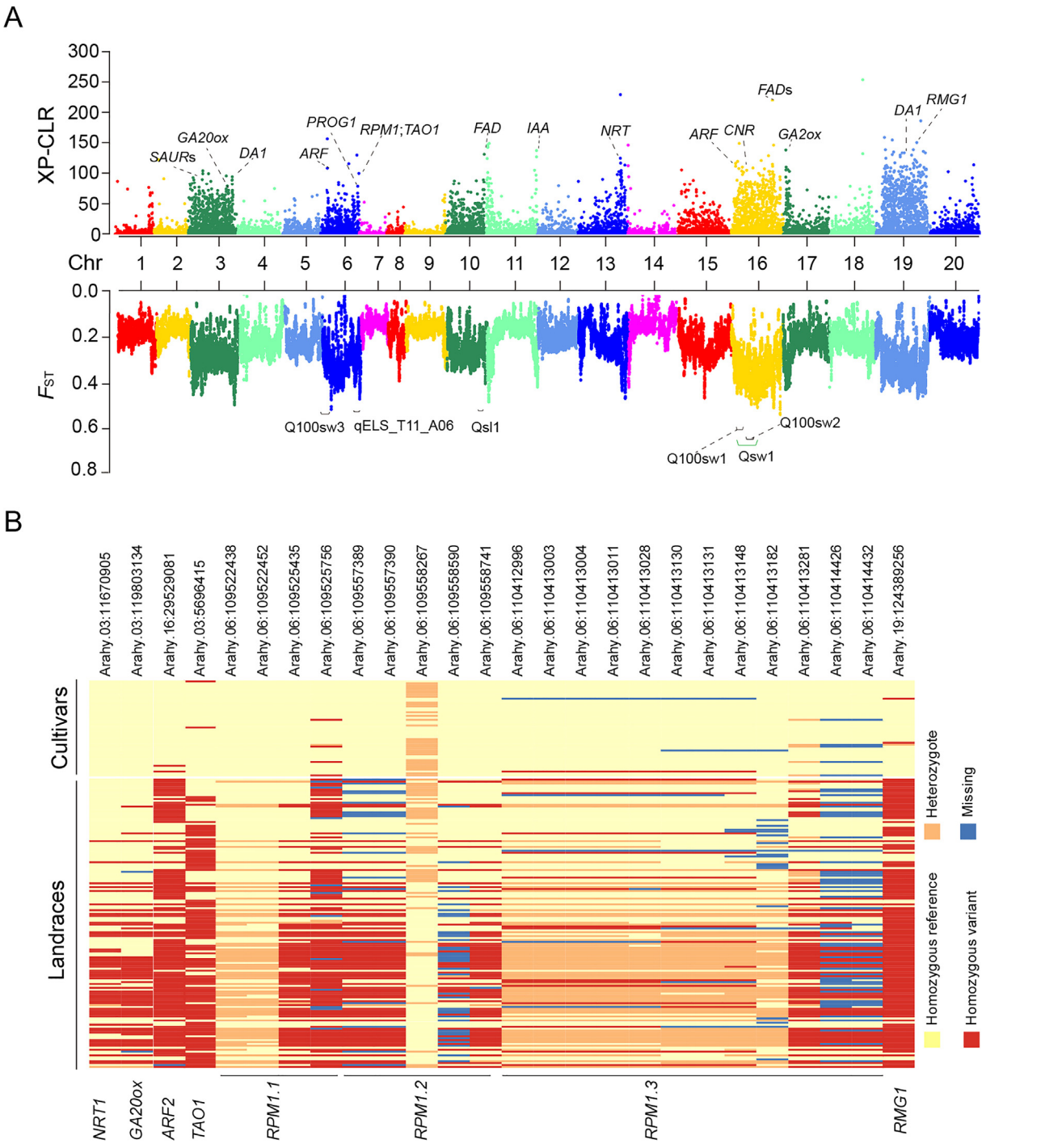


Fig. 3. Selective sweep regions between the improved cultivars and landraces during peanut modern breeding. (A) The genome-wide distribution of the F_{ST} (top panel) and XP-CLR (bottom panel) scores along the chromosomes. The threshold was defined by the top 5% of the F_{ST} and XP-CLR (landrace/ improved cultivar) values. Candidate genes and previously reported QTLs were indicated. *ARF*, Auxin response factor; *CNR1*, cell number regulator 1; *DA* (LARGE IN CHINESE) 1; *FAD*, fatty acid desaturase; *GA20ox*, gibberellin 20-oxidase; *GA2ox*, gibberellin 2-beta-dioxygenase; *IAA*, auxin-induced gene; *NRT1*, nitrate transporter 1; *PROG1*, prostrate growth 1; *RMG1*, resistance methylated gene 1; *RPM1*, resistance to *P. syringae pv. Maculicola* 1; *SAUR*, small auxin up RNA; *TAO1*, the target of AvrB operation 1. Seed weight QTLs: Q100swt1, Q100swt2, and Q100swt3. Seed width QTL: Qsw1. Seed length QTL: Qsl1. Leaf spot disease-resistant QTL: qELS_T11_A06. **(B)** Haplotype differentiation patterns of NRT1, GA20ox, ARF2, RPM1, TAO1, and RMG1. Across the improved cultivars and landraces, significant differentiation patterns of haplotypes were shown within the candidate genes. Geographic maps were generated using R packages.

cultivar breeding. The top gene enrichment gene ontology (GO) terms were "RNA surveillance", "response to auxin", "nitrogen catabolism", and "tyrosine metabolism" (Table S5). Among these genes, 48 are candidate homologs of known plant auxin responding-related genes (e.g., *SAUR*s and *IAA*s). In addition, we identified several genes known to influence the plant yield by comparing the improved cultivars (high-yield) and the landraces (low-yield) (Fig. 3A; Table S6).

Selection pressure analyses revealed several genes associated with the yield, which is consistent with this trait being the major target of peanut breeding. Especially, the selective sweep signals agreed with six previously identified QTLs related to seed size

and weight, including a sweep region from 8.8 to 11.7 Mb on chromosome 16 with two QTLs of seed weight (Q100sw2) and seed width (Qsw1) [36]. Known yield-related genes identified in the sweep regions included four cell number regulator (*CNR*) homolog genes that are responsible for increasing the fruit size in maize and tomato [37,38], two *DA1* homologs that have been proven to play a key role in the maternal control of seed size by regulating cell proliferation in Arabidopsis [39], and one *ARF2* homolog (*Arahy.7E4SNA*) that encodes a seed-size regulator auxin response factor 2 in Arabidopsis [40]. When the haplotype differentiation patterns of *AhARF2* were compared between the improved cultivars and the landraces, a missense SNP variation (c.964 T > C|p.322 Ile > Thr) was identified. Interestingly, 96% of the improved cultivars carried haplotype T, in contrast to the accessions from the landraces, which mainly possessed haplotype C (71%) (Fig. 3B; Table S7).

Genes encoding phytohormones and other genes associated with the plant architectures were also recognized in our analysis. These genes include gibberellin 20 oxidase (GA20ox) and gibberellin 2-beta-dioxygenase (GA2ox), which are involved in plant height regulation [41]. A comparison of the haplotypes between the improved cultivars and the landraces in these genes revealed the presence of a major haplotype for *AhGA20ox* (c.964 G > A|p.322 Glu > Lys), which accounted for 100% in the improved cultivars and had a low allele frequency (64%) in the landraces (Fig. 3B; Table S7). Moreover, *PROG1* (*prostrate growth 1* homolog), which controls a key transition from prostrate to erect growth in rice [42,43], was detected in chromosome A06 sweep regions.

As evident from the relevant literature and our genetic data, the improved cultivars had been bred to achieve near-optimal yields and possess the combined characteristics of high oil content, high efficiency of fertilizer utilization, and high resistance to abiotic and biotic stresses. Pathogen resistance, nitrate uptake, and lipid metabolism-related genes were consistently detected and were considered a high-confidence gene set (Fig. 3A; Table S6). Eight genes related to lipid metabolism were detected in the sweep regions, including a region on chromosome 16 with four fatty acid desaturase genes in tandem. Fourteen genes encoding NRT1/PTR family protein were identified in several selected regions, which alludes to their potential contribution to nutrient utilization during peanut improvement. The haplotype differentiation patterns showed that a missense SNP variation in *AhNRT1* differentiated the improved cultivars from the landraces (Fig. 3B; Table S7). A 29-gene set involved in disease resistance comprised genes encoding the resistance methylated gene 1 (RMG1), the target of AvrB operation (TAO1), and disease resistance protein (RPM1) (Fig. 3A; Table S6). Interestingly, the RPM1 sweep region included a QTL associated with leaf spot disease resistance [44]. Examining the haplotypes within these disease resistance genes revealed that all detected missense SNP variations were significantly differentiated between the improved cultivars and the landraces (Fig. 3B). This finding hints at the contribution of these variations to increased disease resistance in the improved cultivars. Our results suggest that seed size, plant architecture, and disease resistance were the major loci selected during the breeding history of the improved peanut cultivars.

Candidate genes related to seed traits

Seed size is a crucial trait related to crop yield. We employed GWAS to screen for candidate genes related to seed length, seed width, and seed weight regulation. Significant signals were identified in the three phenotypes and two genes were found to play roles in seed weight and seed length, however, no gene was identified in relation to seed width (Figs. 4 and 5; Fig. S3). The study showed a strong association with seed weight on chromosome 16 (8.37–8.55 Mb), which included 8 genes (Fig. 4A and 4B; Table S8). Consistently, the identified genomic regions overlapped with the selective sweep signal (Fig. 3A) and the previously reported QTLs for seed weight [36]. Among them, we noted one nonsynonymous SNP in the sixth exon of the *AhFAX1* (*Arahy.QEH0EE*) locus, which encodes a chloroplast inner envelope localized member of the Tmemb14 gene family involved in fatty acid and lipid homeostasis and probably functions as a fatty acid transporter from the plastid [45]. The nonsynonymous SNP resulted in a cysteine-to-tyrosine substitution in the AhFAX1 TMEM14 domain (Fig. 4C). Gene-based association analysis revealed that haplotype A (TAT, tyrosine) is mainly found in accessions with a higher seed weight, while haplotype G (TGT, cysteine) mainly occurs in accessions with a lower seed weight (Fig. 4D; Table S1). The highest *AhFAX1* expression was detected in the leaf tissue and increased gradually with the seed development (Fig. 4E), which is a critical period for seed oil accumulation. Protein structural modeling revealed that the cysteine-to-tyrosine substitution created an α-helical domain that is absent in the *AhFAX1* (haplotype G) protein (Fig. 4F). We further performed transgenic experiments in Arabidopsis (because of the lack of efficient transgenic methods for peanut) to validate the function of *AhFAX1*. When compared with the wild-type plants, both the *AhFAX1* with haplotype G (*35S:AhFAX1^{hapG}*) and *AhFAX1* with haplotype A (*35S:AhFAX1^{hapA}*) overexpression lines had higher seed size and seed weight. Moreover, the 1000-seed weight of the *35S:AhFAX1^{hapA}* transgenic lines was significantly higher than that of *35S:AhFAX1^{hapG}* lines (Fig. 4G). The total lipid content of the seeds increased by 11.4%, 12.0%, 24.2%, and 19.5% in the overexpressed lines of *35S:AhFAX1^{hapG}*#2, *35S:AhFAX1^{hapG}*#5, *35S:AhFAX1^{hapA}*#3, and *35S:AhFAX1^{hapA}*#8 compared to the WT, respectively (Fig. S4). In addition, an analysis of the fatty acid composition of mature seeds revealed that palmitic acid (C16:0) was significantly increased in the transgenic lines than the WT; however, eicosenic cis (C20:1) was significantly reduced (Fig. S4). These results signify that the cysteine-to-tyrosine substitution in *AhFAX1* may favor total lipid accumulation and change the fatty acid composition to increase the seed weight in peanuts.

Furthermore, we found higher association signals with seed length on chromosome 15 (Fig. 5A; Table S9), in which an SNP of 1,045 bp upstream from the start codon of *Arahy.SV5NHS* was identified (Fig. 5B and C). *Arahy.SV5NHS* (designed as *AhDPB2*, an ortholog of *AtDPB2* in Arabidopsis) encodes a protein similar to DNA polymerase epsilon subunit B, which is required for DNA replication and cell cycle progression [46,47]. The seed length of accessions with haplotype A was significantly higher than that of accessions with haplotype G (Fig. 5D; Table S1). In agreement with this finding, quantitative polymerase chain reaction revealed a markedly elevated expression of *AhDPB2* at the seed 3 stage (50 DAF), which is a seed rapid growth stage (Fig. 5E). Previous work has demonstrated that suppressing the expression of *AtDPB2* results in prolonged cell cycle division and enlarged seed embryos [48]. Since the SNP is located in the promoter region of *AhDPB2*, we speculated that it plays a role in seed size by modulating the expression of *AhDPB2* (e.g., by altering a promoter element). We randomly selected five large-seeded and five small-seeded accessions to explore the expression at the seed 3 stage by using transcriptomic data. The expression level of *AhDPB2* in large-seeded accessions was significantly lower than that in the small-seeded accessions (Fig. 5F). We further validated the *AhDPB2* expression level by RNA-seq in different seed-size accessions (Fig. 5G). These results indicate that *AhDPB2* might play a negative regulatory role in seed length development in peanuts.

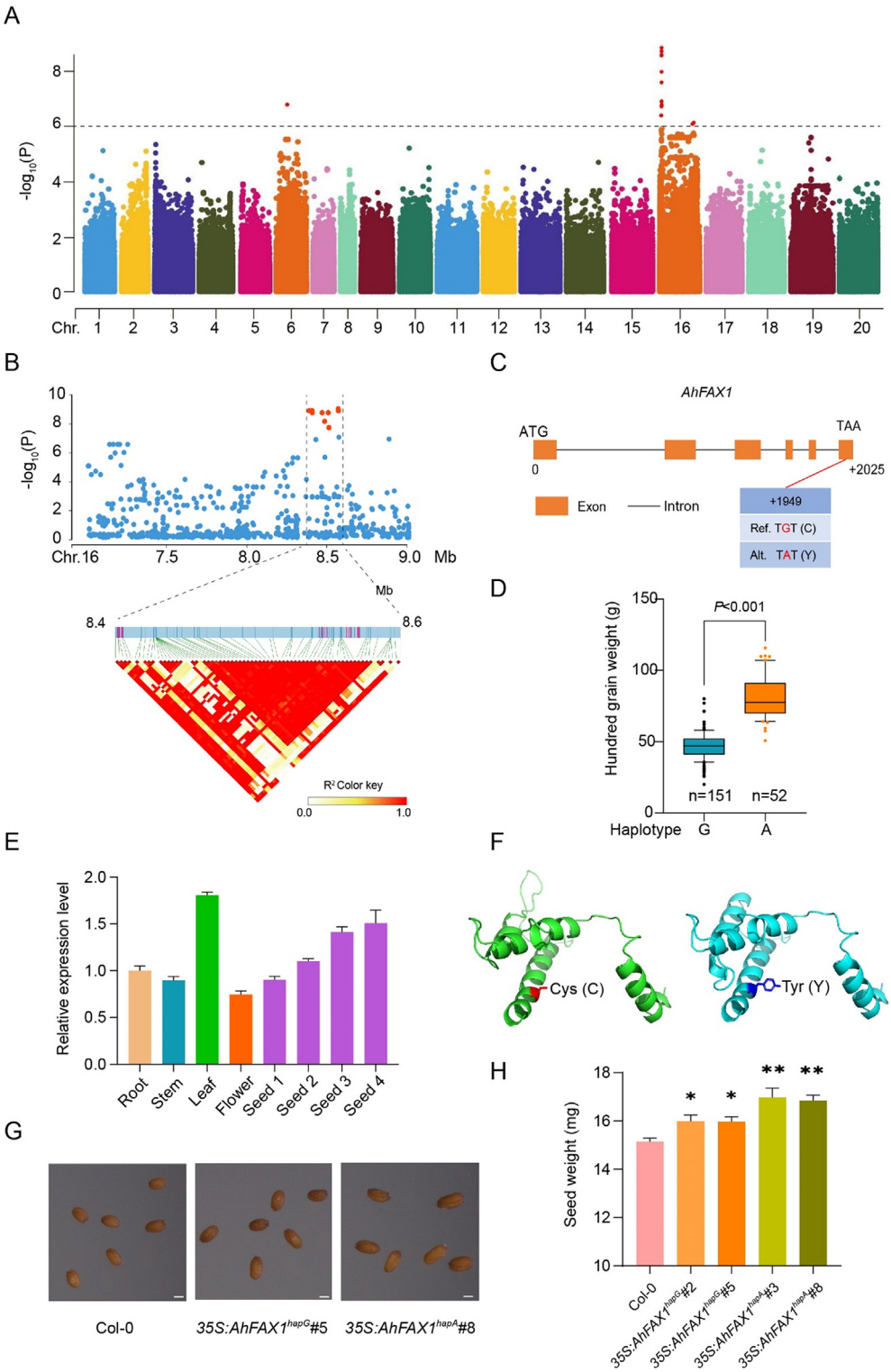


Fig. 4. Identification of the candidate gene *FAX1* for seed weight based on GWAS. (A) The GWAS signals for seed weight in the 20 chromosomes. The horizontal red line represents the significance threshold ($-\log_{10}P > 6$). The red dots indicate strongly associated SNPs. (B) Local Manhattan plot (top panel) on chromosome 16 and LD heat map (bottom panel). The candidate region (8.4–8.6 Mb) lies between the dashed lines. The red dots indicate the strongly associated loci SNP containing the candidate gene *FAX1*. (C) Exon–intron structure of *AhFAX1* in two haplotypes, 'G' and 'A' allele. (D) Box plots for seed weight for the two haplotypes mentioned above (152 vs 51 accessions). Dots show individual data points outside the 10–90 percentile ranges. The $P < 0.001$ indicates a significant difference between the 'G' and 'A' haplotypes by a Student's *t*-test. (E) Relative expression of *AhFAX1* in different tissues, including roots, stems, leaves, flowers, and seeds of developmental stages (seed 1–4) was determined by qRT-PCR. (F) Protein structural modeling of AhFAX1. The protein structure of AhFAX1 (haplotype G, top panel; haplotype A, bottom panel). The site for the Cys/Tyr substitution is marked in red and blue, respectively. (G) The mature seeds of WT, the *AhFAX1* with haplotype 'A' overexpression line (*35S:AhFAX1*^{hapA}#5), and the *AhFAX1* with haplotype G overexpression line (*35S:AhFAX1*^{hapG}#8) in Arabidopsis. Bar = 500 μm. (H) The average 1000 seed weight of two independent lines of *35S:AhFAX1*^{hapG}, *35S:AhFAX1*^{hapA},and Col-0. The data are means ± SD from at least 10 seedlings in three replicates. *, ** indicate significant difference from the wild type plants by using Student's *t*-test with $P < 0.05$ and $P < 0.01$, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

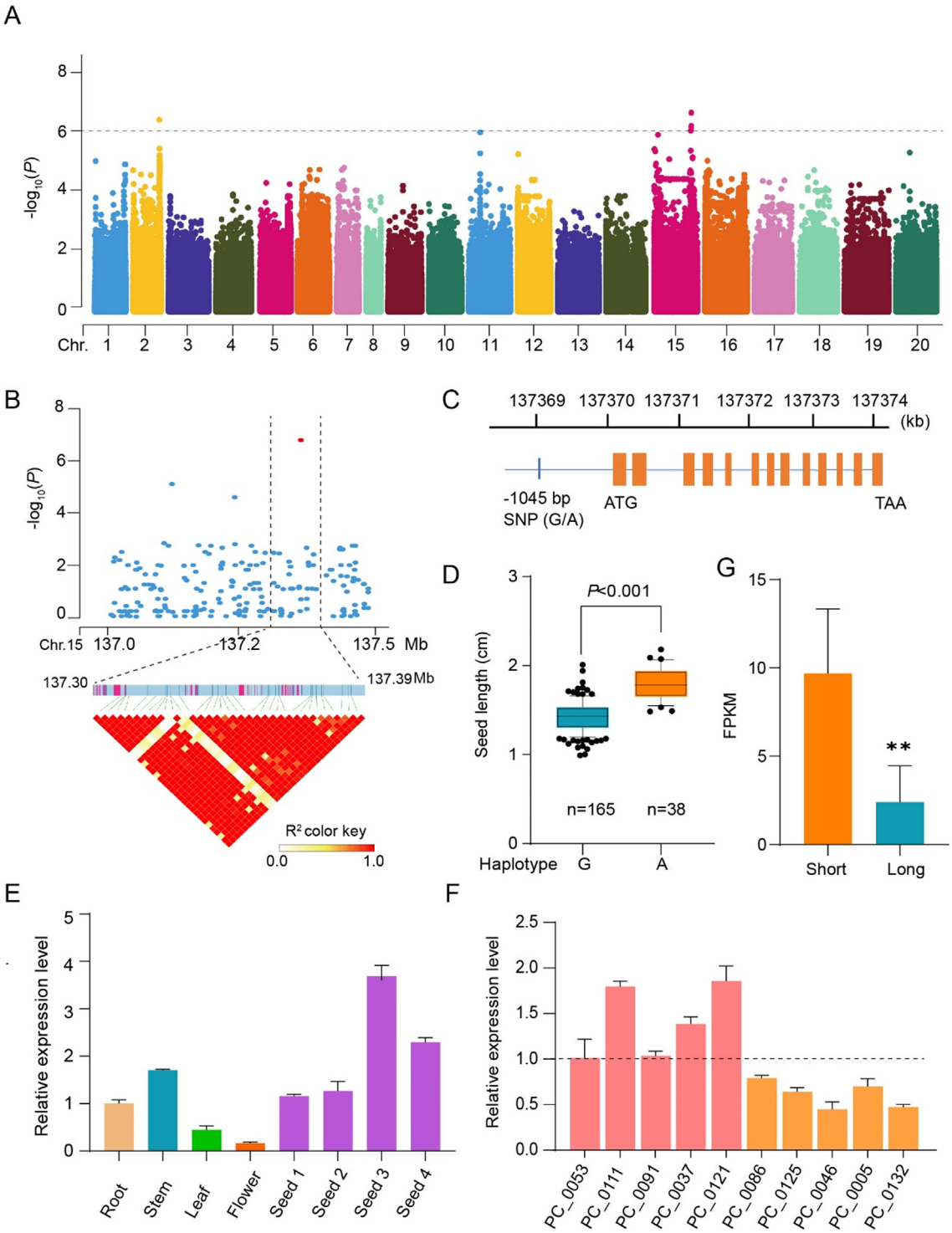


Fig. 5. Identification of candidate gene *AhDPB2* for seed length based on GWAS. (A) Manhattan plots for seed length in the full population. The horizontal dashed line indicates the significance threshold ($P = 1 \times 10^{-6}$). (B) Manhattan plot on chromosome 15 (top panel) and LD heat map (bottom panel) of the candidate region between 137.30 and 137.39 Mb (dashed lines). The red dots indicate the SNPs strongly associated with *AhDPB2*. (C) Gene structure and position of *AhDPB2* on the genome. The polymorphism of 'G' or 'A' haplotypes occurred on the promoter region (-1045). (D) Boxplots for seed length for the 'G' or 'A' haplotypes (165 vs 38 accessions). The dots show individual data points outside the 10–90 percentile ranges. The $P < 0.001$ indicates a significant difference between the 'G' and 'A' haplotypes by a Student's *t*-test. (E) The expression of *AhDPB2* in roots, stems, leaves, flowers, and seeds of different developmental stages (seed 1–4) was determined by qRT-PCR. (F) The expression level of *AhDPB2* in five short-seed and five long-seed accessions at the seed 3 stage was determined by qRT-PCR. The detailed information of the accessions was deposited in Table S1. (G) The transcriptomic level of *AhDPB2* in short and long seeds was plotted with FPKM. ** indicates significant difference with $P < 0.01$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Discussion

The cultivated peanut is categorized into subsp. *fastigiata* and subsp. *hypogaea*, which include six botanical varieties based on their morphologic phenotypes. The accessions resequenced in our collection belong to five botanical varieties that display a high geographic diversity. Seventy-two accessions were collected from South America, which is considered as the primary and secondary center for peanut domestication. The π values implied that the accessions from South America had higher genetic diversity and those from Africa had a lower diversity; however, both exhibited higher diversity than the accessions from Asia. These findings agree with the idea that peanut was introduced to Africa from America by the Portuguese in the 16th century and then spread to China, India, Japan, Malaysia, and elsewhere. These results also confirm that polyploidization, self-pollination, domestication, and selection during breeding are the primary causes for the narrow genetic base of the cultivated peanut [6]. Moreover, the significant proliferation of peanuts 300–500 years ago could be attributed to these factors (Fig. 2A).

Our population structure and PCA suggest that var. *hypogaea* and var. *hirsuta* are intermingled and belong to subsp. *hypogaea*. This result is consistent with a recent targeted genotyping by sequencing study of 320 cultivated peanuts [49]. The var. *hypogaea* has been postulated to represent the most ancient variety owning to its runner habit, the lack of floral spikes and branching patterns, which are the typical features of wild *Arachis* species [50]. However, our data suggest that var. *peruviana* diverged earlier than the other varieties, with high diversity and fixation values. The most ancient macrofossil suggests that the peanut existed 8500 years ago in the Zana valley in Northern Peru [51]. Our demographic data revealed that var. *peruviana* diverged approximately 11,441 years ago, which is in line with the multiple lines of evidence, including genetic analyses. These findings indicate that the peanut originated approximately 10,000 years ago [23].

Selection pressure analysis confirmed that the improved cultivars have been bred to achieve near-optimal yields and exhibit the combined characteristics of high oil content, high efficiency of fertilizer utilization, and high resistance to abiotic and biotic stresses. Moreover, several previously reported QTLs (Qsw1, Q100sw2, and qELS_T11_A06) related to seed size and disease resistance overlapped with the sweep regions, which contain some known key genes (e.g., *ARF2* and *RPM1*) (Fig. 3A). Plant architecture plays a crucial role in the yield, and key genes (*GA2ox*, *GA20ox*, and *PROG1*) associated with the plant architecture were also identified in our analysis. We further identified some selective signals involving novel genes, which may be related to lipid metabolism, fertilizer utilization, and abiotic and biotic stress resistance. Moreover, we identified and validated two novel genes associated with higher seed weight (*AhFAX1*) and seed length (*AhDPB2*) by GWAS (Figs. 4 and 5), which may be important candidate gene resources for accelerating yields in future peanut breeding.

Conclusion

In summary, we performed a population genomics study of 203 peanut accessions from across the world to understand their genetic diversity and selection sweeps and identified candidate genes associated with desirable traits. Our study provides a strong basis for future work on the breeding and genetic aspects of this important oil, food, and forage crop.

CRediT authorship contribution statement

Yiyang Liu: Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Libin Shao:** Formal analysis, Writing – original draft. **Jing Zhou:** Formal analysis, Writing – original draft. **Rongchong Li:** Formal analysis. **Manish K. Pandey:** Formal analysis. **Yan Han:** Resources. **Feng Cui:** Resources. **Jialei Zhang:** Resources. **Feng Guo:** Resources. **Jing Chen:** Formal analysis. **Shihua Shan:** Formal analysis. **Guangyi Fan:** Writing – original draft. **He Zhang:** Formal analysis. **Inge Seim:** Writing – original draft. **Xin Liu:** Writing – review & editing. **Xinguo Li:** Writing – original draft. **Rajeev K. Varshney:** Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Guowei Li:** Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Shubo Wan:** Writing – original draft.

Data availability

All the genomics sequence datasets have been deposited in the NCBI database with BioProject number PRJNA646040 and the CNSA database (CNGB Nucleotide Sequence Archive) with the accession number CNP0001330.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Taishan Scholar Program (No. tsqn20161058), the National Natural Science Foundation of China (31871665), the Programs from department of Science and Technology of Shandong Province (YDZX20203700001861, 2019LZGC017, 2020LZGC001) and the Innovation Program of SAAS (CXGC2021A09), and the open fund of Shandong Provincial Key Laboratory of Plant Stress. We would like to thank MJEditor (www.mjeditor.com) for its linguistic assistance during the preparation of this manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2022.01.016.

References

[1] Toomer OT. Nutritional chemistry of the peanut *(Arachis hypogaea)*. Crit Rev Food Sci 2018;58(17):3042–53.
[2] Wondracek-Lüdke DC, Custodio AR, Simpson CE, Valls JFM. Crossability of Arachis valida and B genome Arachis species. Genet Mol Res 2015;14(4):17574–86.
[3] Krapovickas A, Gregory WC, Williams DE, Simpson CE. Taxonomy of the genus Arachis (Leguminosae). Bonplandia 2007;16:7–205.
[4] Ferguson ME, Bramel PJ, Chandra S. Gene diversity among botanical varieties in peanut (*Arachis hypogaea* L.). Crop Sci 2004;44(5):1847–54.
[5] Xu W, Wu Di, Yang T, Sun C, Wang Z, Han B, et al. Genomic insights into the origin, domestication and genetic basis of agronomic traits of castor bean. Genome Biol 2021;22(1). doi: https://doi.org/10.1186/s13059-021-02333-y.
[6] Mallikarjuna N, Varshney RK. Genetics, genomics and breeding of peanuts. CRC Press; 2014.
[7] Otyama PI, Kulkarni R, Chamberlin K, Ozias-Akins P, Chu Y, Lincoln LM, et al. Genotypic characterization of the US peanut core collection. G3: Genes, Genomes, Genet 2020;10(11):4013–26.
[8] Zou K, Kim KS, Kim K, Kang D, Park YH, Sun H, et al. Genetic diversity and Genome-Wide association study of seed aspect ratio using a High-Density SNP array in peanut (*Arachis hypogaea* L.). Genes (Basel) 2020;12(1):2.
[9] Holbrook CC, Burow MD, Chen CY, Pandey MK, Liu L, Chagoya JC, et al. Recent advances in peanut breeding and genetics. Peanuts 2016:111–45.
[10] Zhao C, Shao C, Yang Z, Wang Y, Zhang X, Wang M. Effects of planting density on pod development and yield of peanuts under the pattern of precision planted peanuts. Legume Res 2017;40(5):901–5.
[11] Yang S, Zhang J, Geng Y, Tang Z, Wang J, Guo F, et al. Transcriptome analysis reveals the mechanism of improving erect-plant-type peanut yield by single-

seeding precision sowing. PeerJ 2021;9:e10616. doi: https://doi.org/10.7717/peerj.10616.
[12] Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G, et al. The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nat Genet 2019;51(5):877–84.
[13] Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, et al. SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience 2018;7(1):1–6.
[14] Kendig KI, Baheti S, Bockol MA, Drucker TM, Hart SN, Heldenbrand JR, et al. Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. Front Genet 2019;10. doi: https://doi.org/10.3389/fgene.2019.00736.
[15] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297–303.
[16] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for Whole-Genome association and Population-Based linkage analyses. The American Journal of Human Genetics 2007;81(3):559–75.
[17] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38(8):904–9.
[18] Subramanian B, Gao S, Lercher MJ, Hu S, Chen W. Evolview v3: A webserver for visualization, annotation, and management of phylogenetic trees. Nucleic Acids Res 2019;47(W1):W270–5.
[19] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19(9):1655–64.
[20] Fan W, Lu J, Pan C, Tan M, Lin Q, Liu W, et al. Sequencing of Chinese castor lines reveals genetic signatures of selection and yield-associated loci. Nat Commun 2019;10(1). doi: https://doi.org/10.1038/s41467-019-11228-3.
[21] Zhang C, Dong S, Xu J, He W, Yang T. PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 2019;35(10):1786–8.
[22] Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet 2017;49(2):303–9.
[23] Bertioli DJ, Abernathy B, Seijo G, Clevenger J, Cannon SB. Evaluating two different models of peanut's origin. Nat Genet 2020;52(6):557–9.
[24] Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EKS, et al. The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. Nat Genet 2016;48(4):438–46.
[25] Excofffier L, Marchi N, Marques DA, Matthey-Doret R, Gouy A, Sousa VC. Fastsimcoal2: Demographic inference under complex evolutionary scenarios. Bioinformatics 2021.
[26] Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Res 2010;20(3):393–402.
[27] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011;27(15):2156–8.
[28] Mi X, Wegenast T, Utz HF, Dhillon BS, Melchinger AE. Best linear unbiased prediction and optimum allocation of test resources in maize breeding with doubled haploids. Theor Appl Genet 2011;123(1):1–10.
[29] Viana JMS, de Almeida ÍF, de Resende MDV, Faria VR, E Silva FF. BLUP for genetic evaluation of plants in non-inbred families of annual crops. Euphytica 2010; 174(1): 31-9.
[30] Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet 2010;42(4):348–54.
[31] Han Y, Li R, Liu Y, Fan S, Wan S, Zhang X, et al. The major intrinsic protein (MIP) family and their function under salt-stress in peanut. Front Genet 2021;12:153.
[32] Xiao Z, Tang F, Zhang L, Li S, Wang S, Huo Q, et al. The Brassica napus fatty acid exporter FAX1-1 contributes to biological yield, seed oil content, and oil quality. Biotechnol Biofuels 2021;14(1). doi: https://doi.org/10.1186/s13068-021-02035-4.
[33] Lin W, Peng Y, Li G, Arora R, Tang Z, Su W, et al. Isolation and functional characterization of PgTIP1, a hormone-autotrophic cells-specific tonoplast aquaporin in ginseng. J Exp Bot 2007;58(5):947–56.
[34] Otyama PI, Wilkey A, Kulkarni R, Assefa T, Chu Ye, Clevenger J, et al. Evaluation of linkage disequilibrium, population structure, and genetic diversity in the US peanut mini core collection. BMC Genomics 2019;20(1). doi: https://doi.org/10.1186/s12864-019-5824-9.
[35] Smartt J. The groundnut crop: A scientific basis for improvement. Springer Science & Business Media; 2012.
[36] Zhang S, Hu X, Miao H, Chu Y, Cui F, Yang W, et al. QTL identification for seed weight and size based on a high-density SLAF-seq genetic map in peanut (Arachis hypogaea L.). Bmc Plant Biol 2019;19(1):1–15.
[37] Alpert KB, Grandillo S, Tanksley SD. Fw 2.2:a major QTL controlling fruit weight is common to both red- and green-fruited tomato species. Theor Appl Genet 1995;91-91(6-7):994–1000.
[38] Guo M, Rupe MA, Dieter JA, Zou J, Spielbauer D, Duncan KE, et al. Cell number regulator1 affects plant and organ size in maize: Implications for crop yield enhancement and heterosis. Plant Cell 2010;22(4):1057–73.
[39] Li Na, Xu R, Li Y. Molecular networks of seed size control in plants. Annu Rev Plant Biol 2019;70(1):435–63.
[40] Vert G, Walcher CL, Chory J, Nemhauser JL. Integration of auxin and brassinosteroid pathways by Auxin Response Factor 2. Proc Natl Acad Sci 2008;105(28):9829–34.
[41] Yamaguchi S. Gibberellin metabolism and its regulation. Annu Rev Plant Biol 2008;59(1):225–51.
[42] Jin J, Huang W, Gao J-P, Yang J, Shi M, Zhu M-Z, et al. Genetic control of rice plant architecture under domestication. Nat Genet 2008;40(11):1365–9.
[43] Tan L, Li X, Liu F, Sun X, Li C, Zhu Z, et al. Control of a key transition from prostrate to erect growth in rice domestication. Nat Genet 2008;40(11):1360–4.
[44] Agarwal G, Clevenger J, Pandey MK, Wang H, Shasidhar Y, Chu Ye, et al. High-density genetic map using whole-genome resequencing for fine mapping and candidate gene discovery for disease resistance in peanut. Plant Biotechnol J 2018;16(11):1954–67.
[45] Li N, Gügel IL, Giavalisco P, Zeisler V, Schreiber L, Soll J, et al. FAX1, a novel membrane protein mediating plastid fatty acid export. Plos Biol 2015;13(2):e1002053.
[46] Pedroza-Garcia JA, Domenichini S, Mazubert C, Bourge M, White C, Hudik E, et al. Role of the Polymerase sub-unit DPB2 in DNA replication, cell cycle regulation and DNA damage response in Arabidopsis. Nucleic Acids Res 2016;44(15):7251–66.
[47] Ronceret A, Guilleminot J, Lincker F, Gadea-Vacas J, Delorme V, Bechtold N, et al. Genetic analysis of two Arabidopsis DNA polymerase epsilon subunits during early embryogenesis. Plant J 2005;44(2):223–36.
[48] Jenik PD, Jurkuta REJ, Barton MK. Interactions between the Cell Cycle and Embryonic Patterning in Arabidopsis Uncovered by a Mutation in DNA Polymerase ε. Plant Cell 2005;17(12):3362–77.
[49] Zheng Z, Sun Z, Fang Y, Qi F, Liu H, Miao L, et al. Genetic diversity, population structure, and botanical variety of 320 global peanut accessions revealed through tunable Genotyping-by-Sequencing. Sci Rep 2018;8(1). doi: https://doi.org/10.1038/s41598-018-32800-9.
[50] Krapovickas A. The origin, variability and spread of the groundnut (Arachis hypogaea). In: The domestication and exploitation of plants and animals. Routledge; 2017. p. 427–42.
[51] Dillehay TD, Rossen J, Andres TC, Williams DE. Preceramic adoption of peanut, squash, and cotton in northern Peru. Science 2007;316(5833):1890–3.