BMC Plant Biology

RESEARCH

Open Access

# Genome-wide association analysis uncovers the genetic architecture of tradeoff between flowering date and yield components in sesame

Idan Sabag[1,2], Gota Morota[2*] and Zvi Peleg[1*]

## Abstract

**Background:** Unrevealing the genetic makeup of crop morpho-agronomic traits is essential for improving yield quality and sustainability. Sesame (*Sesamum indicum* L.) is one of the oldest oil-crops in the world. Despite its economic and agricultural importance, it is an 'orphan crop-plant' that has undergone limited modern selection, and, as a consequence preserved wide genetic diversity. Here we established a new sesame panel (SCHUJI) that contains 184 genotypes representing wide phenotypic variation and is geographically distributed. We harnessed the natural variation of this panel to perform genome-wide association studies for morpho-agronomic traits under the Mediterranean climate conditions.

**Results:** Field-based phenotyping of the SCHUJI panel across two seasons exposed wide phenotypic variation for all traits. Using 20,294 single-nucleotide polymorphism markers, we detected 50 genomic signals associated with these traits. Major genomic region on LG2 was associated with flowering date and yield-related traits, exemplified the key role of the flowering date on productivity.

**Conclusions:** Our results shed light on the genetic architecture of flowering date and its interaction with yield components in sesame and may serve as a basis for future sesame breeding programs in the Mediterranean basin.

**Keywords:** Sesame, GWAS, Flowering date, Yield components

## Background

Sesame (*Sesamum indicum* L.), one of the oldest oil-crops in the world, was domesticated about 5,500 years ago from *Sesamum indicum* subsp. *malabaricum* [1] in the Indian sub-continent. Sesame is an annual diploid (2n = 2x = 26) species, which belongs to the *Sesamum* genus from the *Pedaliaceae* family. Its seeds are comprised of oil (45–60%), proteins (18-25%), carbohydrates (3-25%), and are rich in essential vitamins and mineral-nutrients [2, 3]. The seeds are being used for an array of products in the food (e.g., high-quality oil, tahini paste, and cooking and backing) and pharmaceutical industries [4]. At present, sesame is cultivated mainly in developing countries, with annual seed production of 6.7 million tons (http://www.fao.org/faostat/en/#data/QC). The global shift toward healthier and more nutritional plant-based food products leads to significantly increased demand for sesame seeds and derivative products. However, despite its economical and agricultural importance,

*Correspondence: morota@vt.edu; zvi.peleg@mail.huji.ac.il
[1] The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, The Hebrew University of Jerusalem, 7610001 Rehovot, Israel
[2] Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA

© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

sesame is considered an 'orphan crop-plant' and has been subjected to limited agronomical and scientific research.

Sesame is a short-day erect plant with an indeterminate florescence and simple or branching rigid stem. Its growth period ranges usually from 12 to 18 weeks, with flowering (i.e., the transition from vegetative to reproductive phase) begins about 30-40 (early) to 70-80 (late) days after sowing [5]. This variation in flowering time could affect the crop adaption to specific agro-system conditions. As the blooming period continues until plant maturation, the flowering date plays a crucial role in both plant architecture and yield components. Sesame yield components include the number of plants per unit area, number of branches per plant, number of capsules per leaf axil, seeds per capsule, and seed weight [6]. Complex tradeoffs between yield component traits have been shown to significantly affect the final seed yield [7]. Branching habit and number of capsules per leaf axil were shown to support higher seed yield [8], whereas dwarf mutants (i.e., small plant height) positively affect seed weight [9].

The advent of next-generation sequencing and genotyping by sequencing (GBS) technologies has provided a means for examining genetic diversity and population structure of crop-plants, which can facilitate the genetic dissection of agronomic traits and integrate them in breeding programs. Genome-wide association studies (GWAS) are a promising approach that connects phenotypic variation and genomic data (i.e., genetic markers) to detect genomic regions underlying complex traits [10]. GWAS were applied successfully for various crop-plants, such as bread wheat (*Triticum aestivum* L. [11]), maize (*Zea mays* L. [12]), rice (*Oryza sativa* L. [13]), and soybean (*Glycine max* L. [14]). In sesame, the GWAS approach was applied to the identification of genomic regions associated with response to biotic [15] and abiotic [16, 17] stresses, quality traits [18, 19], and yield-related traits [20]. The relatively small genome size (~ 375 Mbps), the recent development of genomic resources, and its rich genetic diversity make sesame an ideal model crop for genetic investigation [21].

Here we harness the natural variation among geographically distributed sesame germplasm to underpin the genetic architecture of morpho-agronomic and yield-related traits. Our working hypothesis was that as a consequence of minimal artificial selection processes (associated with modern breeding), sesame preserved rich genetic and phenotypic diversity that will enable detection of novel genomic regions conferring agronomical important traits. The aims of the current study were to (***i***) characterize the genetic diversity in the newly established sesame collection, (***ii***) detect genomic regions contributing to the phenotypic performance, and (***iii***) infer the genetic associations between traits. Our findings shed new light on the interaction between flowering date, morpho-physiological traits, and yield components in sesame.

## Methods

### Plant material and experimental design

A panel of 184 sesame genotypes (most of them are landraces) was assembled from the wide Hebrew University of Jerusalem sesame collection (SCHUJI panel, hereafter), based on their geographical origin and phenotypic observations to capture most of the gene-pool genetic diversity (Supplemental Table S1). These genotypes were grown in a closed net-house over three generations to ensure homozygosity and uniformity. The SCHUJI panel was characterized over two growing seasons (2018 and 2020) at the experimental farm of Hebrew University of Jerusalem in Rehovot, Israel (34°47′N, 31°54′E; 54 m above sea level). The soil at this location is brown-red degrading sandy loam (Rhodoxeralf) composed of 76% sand, 8% silt, and 16% clay. In the 2018 growing season, a complete random factorial (genotypes) block design with seven replicates was employed. Each block consisted of 184 plots sown as a single row, with six plants, 15 cm apart (1-m-long plot). The two plants at the edges of each plot served as borders. The remaining four plants were used for phenotypic characterization and at the end of the experiment, they were harvested to estimate yield components. In the 2020 growing season, the same experimental design was employed with five replicates per genotype. The plot size was 2.6 m × 0.8 m (15 cm between plants) with three rows per plot. Five representing plants from the middle row were used for phenotypic evaluation at maturity and harvested to obtain yield components. In both seasons, the field was treated with fungicides and pesticides to avoid the development of fungal pathogens or insect pests and was weeded manually once a week.

### Phenotypic measurements

Phenotypes were recorded during the whole sesame growing season for each plot. Flowering date (FD) was evaluated visually when 50% of the plants in each plot had at least one open flower. Height to the first capsule (HTFC) and plant height (PH) were measured at maturity from the soil surface to the first capsule and the plant tip, respectively. The reproductive zone of the main stem (RZ) was calculated as the delta between PH and HTFC, and the reproductive index (RI) was calculated as the ratio between RZ and PH (RZ/PH). Before harvest, the number of branches per plant (NBPP) were counted as an average of all individual plants in a plot. At physiological maturity, three plants (2018) and five plants (2020) from

each plot were harvested, and sun-dried. The samples were threshed using the laboratory threshing machine (LD 350, WinterSteiger, Reid, Austria). Seeds were counted using the seeds counting machine (Data Count S25, Data Technologies) and weighted in analytical lab weight to obtain seeds number per plant (SNPP), seed-yield per plant (SYPP), and thousand-seed weight (TSW) for each plot.

#### Statistical analysis of phenotypic data

The JMP ver.15 pro statistical package (SAS Institute, Cary, NC, USA) and R [22] were used for all statistical analyses with a significant threshold of 5%. First, we calculated the best linear unbiased estimate (BLUE) for every trait for each genotype per year and for both years using the lme4 R package [23]. The mixed linear model for BLUE per year was fitted according to the formula:

$$y_{io} = \mu + g_i + b_o + \epsilon,$$

where $y_{io}$ is the phenotypic observation for the $i$th genotype in the $o$th block, $\mu$ is the intercept, $g_i$ is the genotype fixed effect, $b_o$ is the block random effect, and $\epsilon$ is the model residuals.

The BLUE for the combined data from the 2 years was calculated according to the formula:

$$y_{iko} = \mu + g_i + e_k + (ge)_{ik} + b_{ko} + \epsilon,$$

where $y_{iko}$ is the phenotypic observation for the $i$th genotype in the $k$ th year and the $o$th block, $\mu$ is the intercept, $g_i$ is the genotype fixed effect, $e_k$ is the random effect of year, $(ge)_{ik}$ is the random effect of genotype-by-year interaction, $b_{ko}$ is the random effect of block nested within the year, and $\epsilon$ is the model residual. For the calculation of heritability, we fitted the same mixed model as above, except for genotype, which was considered as a random effect. A broad-sense heritability was calculated on the entry-mean basis according to Schmidt et al. [24] using the estimated variance components:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{G\times Y}^2}{n_Y} + \frac{\sigma_\epsilon^2}{n_Y n_r}},$$

where $\sigma_G^2$, $\sigma_{G\times Y}^2$ and $\sigma_\epsilon^2$ are the genetic, genotype-by-year interaction, and residual variances, respectively, and $n_Y$ is the number of years (2) and $n_r$ is the average number of replicates across years (6). The significance of variance components was evaluated by the likelihood ratio test using the lmerTest R package [25].

We performed GWAS for each data set (referred to as 2018 data, 2020 data, and combined data, hereafter). The combined data set was used to infer phenotypic and genomic correlations and to conduct principal components analysis (PCA), k-means clustering, and genomic heritability estimation. K-means clustering analysis was applied using the factoextra R package [26] on the centered and scaled values of each trait. Both PCA and clusters plot were drawn in the JMP ver.15 pro statistical package. Density plots and a correlation matrix were constructed using the ggplot2 [27] and the corrplot [28] R packages.

#### Genotyping and preparation of markers data set

Genomic DNA was extracted from young leaf tissues with a modified CTAB method [29]. We generated GBS data using the procedure described in Elshire et al. [30], with minor changes: 100 ng of genomic DNA and 3.6 ng of total adapters were used. Genomic DNA was restricted with ApeKI enzyme, and the library was amplified with 18 PCR cycles. We used the Zhongzhi No. 13 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000512975.1) reference genome to perform reference-based single nucleotide polymorphisms marker (SNP) calls based on the STACKS 2.3 pipeline (http://catchenlab.life.illinois.edu/stacks). Approximately 4,300,000 reads were obtained, resulting in the detection of 90,542 SNPs. The overall mean depth per site was five, and the proportion of heterozygous was 2%. These SNPs included sites of unknown scaffolds and chloroplast genome. Markers of unknown scaffolds and chloroplast genome were eliminated and all the markers were filtered to a depth quality of 3 using TASSEL ver. 5.0 [31]. Imputation of missing genotypes was performed by BEAGLE ver. 5 [32]. We also removed markers that were tightly linked ($r^2 = 0.99$) using PLINK [33]. Polymorphic sites with < 5% minor allele frequency and > 20% heterozygosity were filtered out by PLINK and TASSEL, respectively. The remaining 20,294 SNPs were used for further analysis.

#### Population structure, kinship, and linkage disequilibrium

We used PCA and a centered identity-by-state matrix (G) constructed from TASSEL to infer population structure in the sesame panel. The ADMIXTURE software [34] was used to estimate a Q-matrix, which is the ancestry among the accessions. To select the number of subpopulations (K), we ran the software from K = 1 to 10 with five-fold cross-validation. This analysis outputs the cross-validation error (%) for a given K. The number of subpopulations was determined as K that produced the lowest cross-validation error. The lowest value of cross-validation error was achieved with K = 7 (Supplemental Table S2), and the Q-matrix was constructed with 1000

bootstrapping. The results from the PCA analysis and ADMIXTURE software outputs were plotted using the ggplot2 R package. The fixation index ($F_{st}$) among the seven subpopulations was calculated with VCFtools [35] according to Weir and Cockerham [36]. Genome-wide linkage disequilibrium (LD) was obtained through pairwise correlations between markers with a sliding window of 10 markers using PLINK [33]. The correlation between markers ($r^2$) was plotted against their physical positions (i.e., base pairs) and the extent of LD pattern and decay was obtained by fitting a non-linear model according to Hill and Weir [37] as described in Marroni et al. [38]. We used a value (in base pairs) in which LD halves from its initial value for defining haplotypes and mining for candidate genes (CG) around significant markers.

#### Genomic heritability and genomic correlations

Genomic heritability estimates and genomic correlations were inferred using the BGLR R package [39]. To obtain genomic heritability for each trait, we fitted a Bayesian univariate genomic best linear unbiased prediction using the equation:

$$\mathbf{Y} = \mu + \mathbf{Xb} + \mathbf{u} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is the vector of single-trait BLUE phenotypes (combined data), $\mu$ is the intercept, $\mathbf{X}$ is a design matrix for fixed effects, $\mathbf{b}$ is the vector of fixed effects containing three PCs to account for population structure, $\mathbf{u}$ is the vector of random effects, and $\boldsymbol{\epsilon}$ is a vector of the model residuals. The following distributions were assumed for random effects:

$$\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{G}\sigma_{\mathrm{u}}^2 & 0 \\ 0 & \mathbf{I}\sigma_{\epsilon}^2 \end{pmatrix}\right],$$

where $\mathbf{G}$ represents the first genomic relationship matrix of VanRaden [40], $\mathbf{I}$ is the identity matrix, and $\sigma_{\mathrm{u}}^2$ is the additive genomic variance explained by genetic markers, and $\sigma_{\epsilon}^2$ is the model residuals. Genomic heritability was calculated as: $h^2 = \frac{\sigma_{\mathrm{u}}^2}{\sigma_{\mathrm{u}}^2+\sigma_{\epsilon}^2}$.

Genomic correlations were estimated using the multivariate version of the model described above, where $\mathbf{Y}$ is the vector of multi-trait phenotypes. The following distributions were assumed for random effects:

$$\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{\mu} \bigotimes \mathbf{G} & 0 \\ 0 & \Sigma_{\epsilon} \bigotimes \mathbf{I} \end{pmatrix}\right],$$

where $\Sigma_{\mu}$ and $\Sigma\epsilon$ refer to the genetic and residual variance-covariance matrices, respectively, and $\otimes$ is the Kronecker product.

Genomic correlations between traits were derived from the genomic variance-covariance matrix as:

$$\mathrm{corr}(\mathbf{u}_1, \mathbf{u}_2) = \frac{\sigma_{\mathbf{u}_1\mathbf{u}_2}^2}{\sqrt{\sigma_{\mathbf{u}_1}^2}\sqrt{\sigma_{\mathbf{u}_2}^2}},$$

where $\mathbf{u}_1$ and $\mathbf{u}_2$ are the breeding values of traits $\mathbf{Y}_1$ and $\mathbf{Y}_2$, $\sigma_{\mathbf{u}_1\mathbf{u}_2}^2$ is the additive genomic covariance between $\mathbf{u}_1$ and $\mathbf{u}_2$, and $\sigma_{\mathbf{u}_1}^2$ and $\sigma_{\mathbf{u}_2}^2$ are additive genomic variances for $\mathbf{Y}_1$ and $\mathbf{Y}_2$, respectively.

#### Association mapping

To identify genomic regions associated with the traits of interest, we used a mixed linear model of Henderson [41] coupled with the first three PCs and G matrices to account for population structure and relatedness among individuals, respectively, using the rrBLUP R package [42]. We fit an additive single-marker GWAS model as the following:

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is the vector of phenotypes, $\mathbf{b}$ is the vector of fixed effects including a SNP and 3 PCs, $\mathbf{u}$ is a vector of random additive genetic effects with mean zero and variance-covariance $\boldsymbol{G}\sigma_u^2$, $\mathbf{X}$ and $\mathbf{Z}$ are the known incidence matrices, and $\boldsymbol{\epsilon}$ is the vector of residuals [43]. The *P*-value threshold ($1.551509 \times 10^{-5}$) was determined by calculating the number of effective independent tests (Meff) as described in Li and Ji [44] with the following formula: $P = 1$ - (1-0.05)^(1/Meff), where $P$ is the genome-wide *P*-value threshold and 0.05 is the desired level of significance, and Meff = 3306. This analysis was conducted on a single-year basis (2018 and 2020) and the BLUEs of the combined two-year data.

#### Haplotypes estimation and candidate gene analysis

Haplotypes encompassing associated markers that corresponded to the LD-decay pattern in the SCHUJI panel were constructed in TASSEL. We included only haplotypes with a frequency greater than 5% and only genotypes that were homozygous in all the genetic markers within a haplotype. The effects of haplotypes on the phenotypic variation were estimated using analysis of variance (JMP pro ver. 15) at a significant level of 5%. The phenotypic response was the BLUE of 2018 and 2020 growing seasons. A haplotype plot was produced using the Raincloud R package [45]. We used LD-decay for mining CGs around significant SNPs and

analyzed them according to the Zhongzhi No. 13 reference genome (https://www.ncbi.nlm.nih.gov/assembly/GCF_000512975.1).

## Results

### High phenotypic diversity among the sesame panel

To test the level of phenotypic diversity among the newly established SCHUJI panel, we characterized the sesame genotypes under Mediterranean basin conditions over two seasons. In general, the SCHUJI panel exhibited rich variation for phenological, plant architecture, and yield components (Fig. 1; Supplemental Table S3; Fig. S1). Flowering date (FD) showed a similar pattern across years and spread along most of the growing season, ranging between 39.86 to 76 and 39.2 to 70 days after sowing (DAS) for 2018 and 2020, respectively. In both years (2018 and 2020), most genotypes (135 and 146) flowered before 55 DAS (Fig. 1A). Plant architecture traits showed a wider range in the 2020 than 2018 seasons for HTFC (29-140 vs. 36-198 cm; Fig. 1B), PH (80-177 vs. 103-251 cm; Fig. 1v), and RZ (20 to 104 vs. 28 to 141 cm; Fig. S1). In contrast, RI exhibited a similar range in both seasons (0.13-0.73 vs. 0.12-0.76; Fig. 1D). Most yield components showed a similar pattern across years (SYPP and TSW) except SNPP which exhibited a much wider variation in 2020 (21-7719 vs. 508-13,591) (Fig. 1F-H).

To test the effect of genotype (G), year (Y), and genotype-by-year interaction (G × Y), we estimated the variance component of each parameter for all the traits. In general, genotype and year had significant effects for most of the traits, with a significant G × Y interaction. The estimates of broad-sense heritability ranged from moderate (0.66) for SNPP to high (0.97) for FD (Table 1).

**Table 1** Variance component estimates of the random effects for each trait

| Trait | Variance component | | | | |
|---|---|---|---|---|---|
| | $\sigma^2_G$ | $\sigma^2_Y$ | $\sigma^2_{G\times Y}$ | $\sigma^2_\epsilon$ | $H^2$ |
| FD | 61.48*** | 0.39$^{n.s.}$ | 2.77*** | 8.84 | 0.97 |
| HTFC | 1069 *** | 214.60*** | 71.49*** | 125.95 | 0.96 |
| PH | 441.81*** | 1174.8*** | 44.18*** | 289.12 | 0.89 |
| RZ | 356.97*** | 394.51*** | 31.97*** | 204.56 | 0.91 |
| RI | 0.02*** | 0.0003* | 0.0009*** | 0.003 | 0.97 |
| NBPP | 4.04*** | 0.72** | 0.64*** | 2.83 | 0.88 |
| SYPP | 28.36*** | 0.72$^{n.s.}$ | 10.21*** | 54.66 | 0.74 |
| SNPP | 1,723,700.70*** | 3,696,689.40* | 854,375.30*** | 5,260,739 | 0.66 |
| TSW | 0.17*** | 0.02*** | 0.03*** | 0.05 | 0.88 |

$\sigma^2_G$ is the genetic variance, $\sigma^2_Y$ is the year variance, $\sigma^2_{G\times Y}$ is the genotype-by-year variance, $\sigma^2_\epsilon$ is the error variance, and $H^2$ is the broad-sense heritability. Flowering date (FD), height to the first capsule (HTFC), plant height (PH), number of branches per plant (NBPP), reproductive index (RI), seed-yield per plant (SYPP), seed number per plant (SNPP), and thousand-seed weight (TSW). n.s., *, and *** indicate non-significant, significant differences at $P \leq 0.05$, and $P < 0.001$, respectively, as determined by the likelihood ratio test

To examine the relationship between vegetative-related traits (late FD, PH, and NBPP) and yield-related traits (RI and yield components), we applied PCA to BLUEs from the combined data. Based on this analysis, three PCs (eigenvalues >1) accounted for 75.1% of the total phenotypic variance among the genotypes (Fig. 2A).

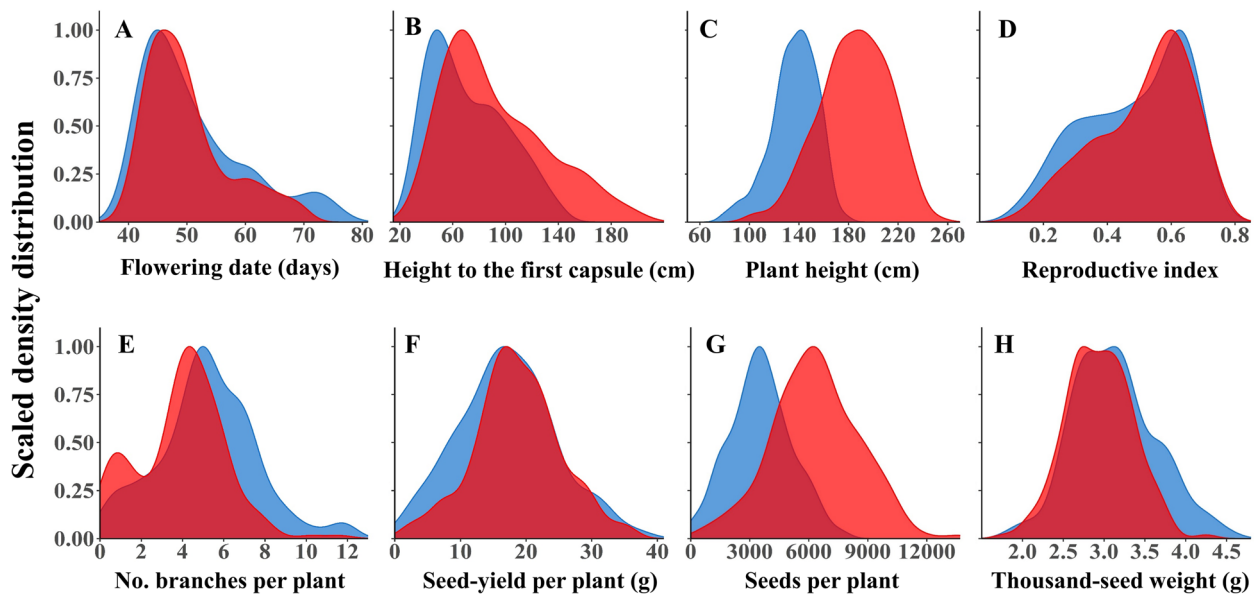


**Fig. 1** Density distribution of the measured traits. *Phenological traits:* (**A**) Flowering date. *Plant architecture traits:* (**B**) height to the first capsule, (**C**) plant height, (**D**) reproductive index, and (**E**) number of branches per plant. *Yield components:* (**F**) seed yield per plant, (**G**) seed number per plant, and (**H**) thousand-seed weight, grown in 2018 (blue) and 2020 (red) seasons

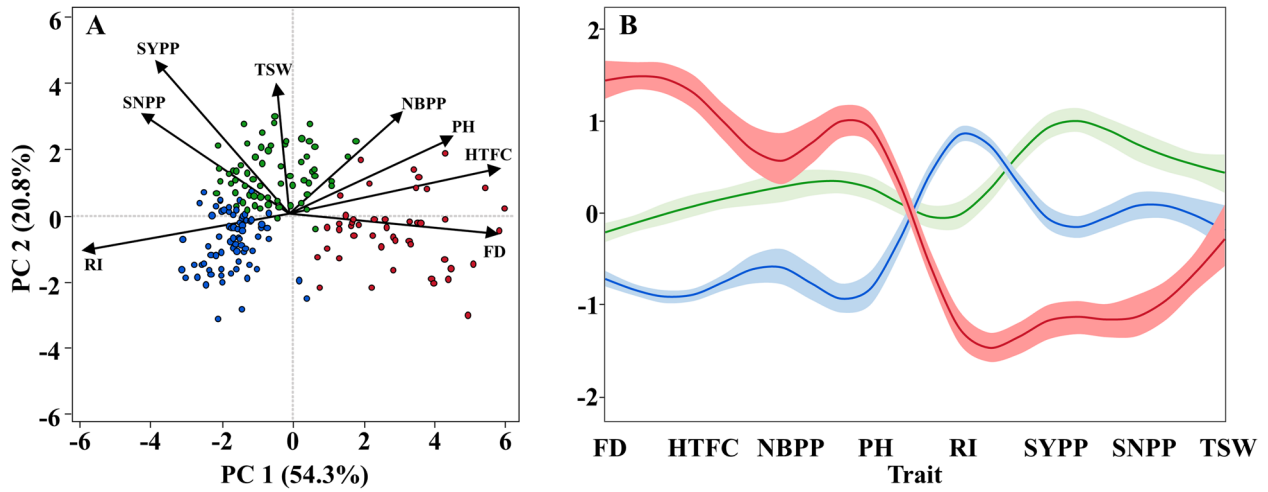


**Fig. 2** Multivariate analysis of the measured traits. (**A**) Principal component (PC) analysis of phenotypic traits of the combined data (2018 and 2020). Each dot represents one genotype. (**B**) K-means clustering analysis of the primary traits. Y-axis is the centered and scaled values for each trait and the lines are the cluster means with their confidence intervals. Clusters 1 (Green), 2, (Blue), and 3 (red) were estimated using the K-means clustering analysis. The traits included Flowering date (FD), height to the first capsule (HTFC), plant height (PH), number of branches per plant (NBPP), reproductive index (RI), seed-yield per plant (SYPP), seed number per plant (SNPP), and thousand-seed weight (TSW)

PC1 explained 54.3% of the variation, loaded positively to FD, HTFC, PH, and NBPP, and negatively loaded to RI, SNPP, SYPP, and TSW. PC2 explained 20.8% of the variation and loaded positively to HTFC, PH, NBPP, SYPP, SNPP, and TSW and negatively loaded to FD and RI. SYPP was positively correlated with SNPP ($r=0.86$) and TSW ($r=0.37$), however, no significant relationship was observed between these two traits ($r=-0.04$; Supplemental Table S4). Overall, FD and PH were positively correlated ($r=0.7$) and both were negatively correlated with RI ($r=-0.89$ and $r=-0.68$, respectively; Fig. 2A; Supplemental Table S4).

### Flowering date affects the final seed-yield via morphological modifications

To further dissect the relationship between flowering date, plant architecture, and yield components, we applied k-means clustering analysis using the combined data. This analysis partitioned the panel into three clusters: early flowering (average 44.84 DAS), mid flowering (48.76 DAS), and late flowering (61.65 DAS) (Fig. 2B; Supplemental Table S5). Comparison between clusters 2 and 1 (early and mid-flowering) shows that they are different in terms of morphological and yield components. Although cluster 2 genotypes have larger RI (0.63 vs. 0.5), genotypes from cluster 1 had greater PH and NBPP (142.23 vs. 166.69 cm and 3.29 vs. 5.12 branches per plant, respectively). These alternations in the plant architecture traits affect the outcome, as we observed higher SYPP, SNPP, and TSW for cluster 1 genotypes (Fig. 2B; Supplemental Table S5). When we compared clusters 1 and 2 to cluster 3, we observed a major difference in FD (61.65 days) that led to a long vegetative phase and higher PH (181.77 cm) and NBPP (5.67), but these cluster 3 genotypes had lower RI (0.31) and lower SNPP, SYPP, and TSW performance.

### Allelic diversity and population structure of the sesame panel

To examine the genetic diversity at the genomic level, we used 20,294 SNP markers that spread along the sesame genome. Principal component analysis on the SNP data explained 30.8% of the genetic variation between genotypes. The PCA did not show any separation between genotypes relative to their geographical origins (Fig. 3A). ADMIXTURE analysis partitioned the panel into four major (K1, K5, K6, and K7) and three minor (K2, K3, and K4) sub-populations (Fig. 3B). $F_{st}$ values of these seven sub-populations ranged from 0.08 to 0.32, which showed a weak to moderate differentiation among the subpopulations (Supplemental Table S6). Genome-wide LD analysis showed that LD decayed rapidly to half of its initial value (0.22) at 58,774 base pairs (Supplemental Fig. S2).

### Genomic heritability and genomic correlations

To elucidate the trait similarity at the genomic level, we computed genomic heritability for each trait and genomic correlations between the measured traits. Estimates of genomic heritability ranged from 0.37 (PH) to 0.58 (SYPP) presenting moderate values for all the traits (Supplemental Table S7). Figure 4 presents phenotypic

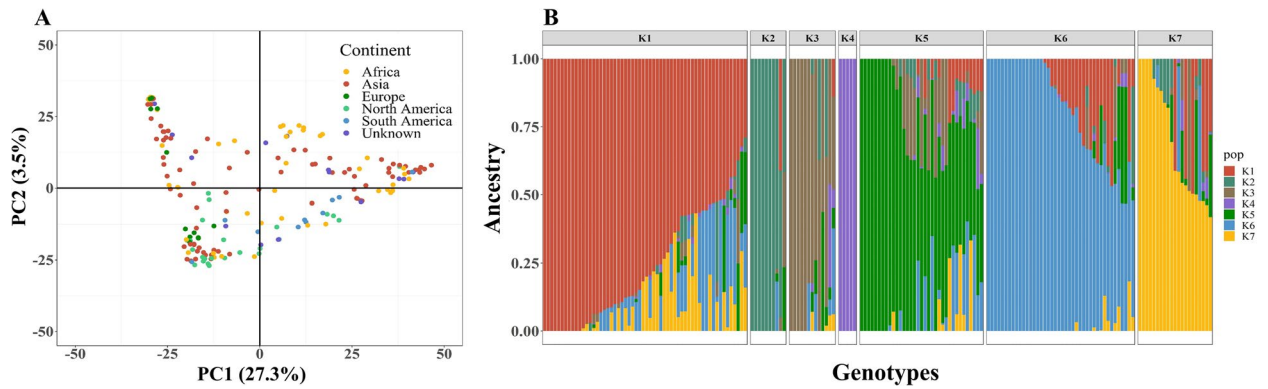


**Fig. 3** Population structure of the SCHUJI panel. (**A**) Principal component analysis and (**B**) ADMIXTURE when K = 7. Every single dot or line represents an individual genotype

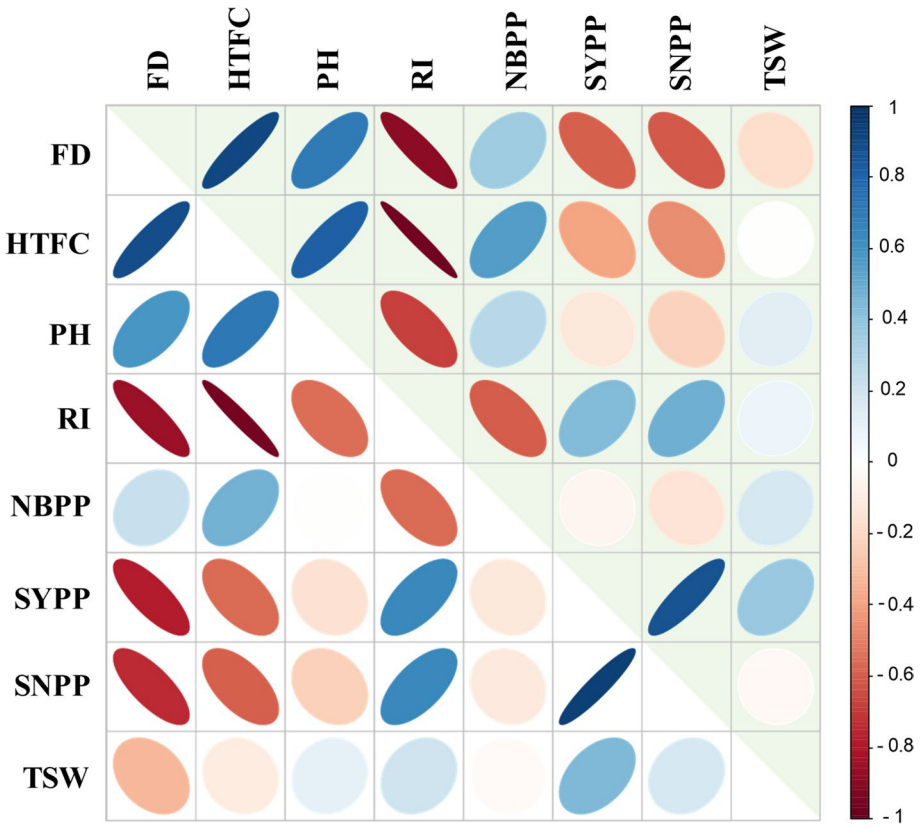


**Fig. 4** Heatmap of phenotypic (upper triangular elements) and genomic (lower triangular elements) correlation matrix between the primary traits: Flowering date (FD), height to the first capsule (HTFC), plant height (PH), reproductive index (RI), number of branches per plant (NBPP), seed-yield per plant (SYPP), seed number per plant (SNPP), and thousand-seed weight (TSW). Colors indicate the level of correlations (r) from positive (blue) to negative (red)

(upper triangular elements) and genomic (lower triangular elements) correlations between the traits. Overall, the phenotypic and genomic correlations showed similar patterns (Fig. 4; supplemental Tables S4, S8). FD was positively correlated with HTFC and PH (0.88 and 0.58, respectively), while negatively correlated with morphological yield-related traits, such as RZ and RI. HTFC was found to be positively correlated with PH (0.71) while

negativity correlated with RZ (−0.75), RI (−0.96), SYPP (−0.56), and SNPP (−0.6). SYPP was positively correlated with SNPP (0.93) and TSW (0.44), whereas SNPP and TSW seemed to be less related (0.17).

### Association mapping of agronomic traits

In an attempt to identify consistent and specific QTL across years, we performed a single-marker regression GWAS for all the traits and detected 11, 19, and 20 SNPs that are associated with trait mean differences for 2018, 2020, and combined data, respectively (Supplemental Table S9). For FD, we identified two major genomic regions on linkage group (LG) 2 and 11 that were consistent across years (4, 4, and 5 SNPs for 2018, 2020, and combined data, respectively; Fig. 5A). In total, 9 SNPs were found significantly associated with the plant architecture traits, including 4 for HTFC (Fig. 5B), 3 for RI (Fig. 5C), and 2 for RZ (Supplemental Fig. S3A). For PH and NBPP, there was no clear peak (Supplemental Fig. S3B-C).

For SYPP and SNPP, we detected one major genomic region on LG2 (except for SNPP in 2018) (Fig. 6A-B). SYPP was associated with 20 SNPs while SNPP was associated with 7 SNPs (Supplemental Table S9). For TSW, we found one SNPs on LG1 that was slightly below the significant threshold ($-\log_{10}(p)=4.77$), only in the 2020 season (Fig. 6C, and Supplemental Table S9). To elucidate the genetic architectures of the measured traits, we compared the mapping results to find SNPs that overlapping across traits. HTFC had one genomic region overlapping with RI on LG16 (Supplemental Table S9). FD had a major genomic region overlapping with SYPP and SNPP (Figs. 5-6; Supplemental Table S9).

### Flowering date promotes yield stability in sesame

A significant genomic region in the length of 96,491 base pairs that contained 10 SNPs on LG2 was found to be associated with FD, SNPP, and SYPP (Supplemental Table S9). In total, four SNPs inside this region overlapped with FD and SYPP and were within the range of the LD-decay pattern of our panel (Supplemental Fig. S2). To explore their influence on the phenotypic variation of these two traits across the two growing seasons, we defined haplotypes for these four overlapping SNPs. Two possible haplotypes were found. The first haplotype (Hap1) was more common ($n=140$) than the second haplotype (Hap2, $n=30$) (Fig. 7). Hap1 was found associated with earliness while genotypes that included Hap2 exhibited late flowering under the Mediterranean climate ($P<0.0001$, Fig. 7A). Moreover, these two haplotypes also differed in yield performance, with Hap1 promoting higher seed-yield (Fig. 7B). To test the phenotypic stability of the haplotypes across years, we perform an analysis of variance for haplotypes, year, and their interaction using FD and SYPP from BLUE per year analysis. While the haplotypes had a significant effect on the traits ($P<0.0001$ for FD and SYPP), we did not observe any significant interaction between haplotypes and year ($P=0.087$ for FD and $P=0.82$ for SYPP). These results may indicate that this genomic region promotes yield stability via modifications in FD. Moreover, we obtained differences in

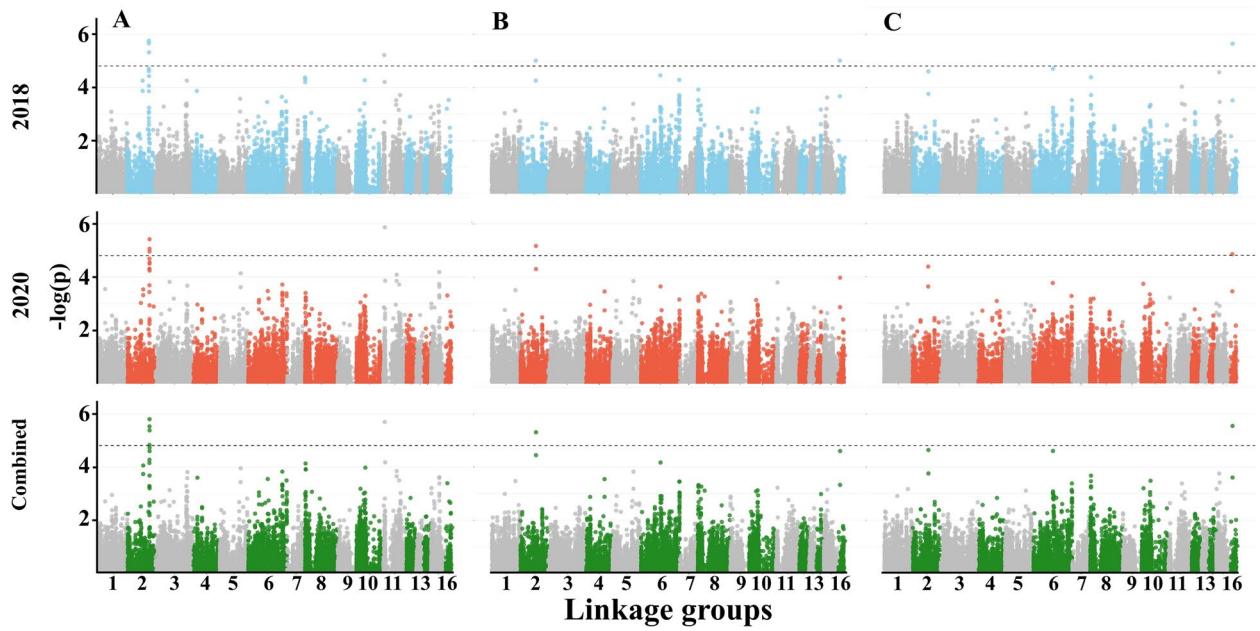


**Fig. 5** Manhattan plots for phenological and plant architecture traits. (**A**) Flowering date, (**B**) height to the first capsule, and (**C**) reproductive index in 2018, 2020, and combined data. The dashed line represents the genome-wide significance threshold

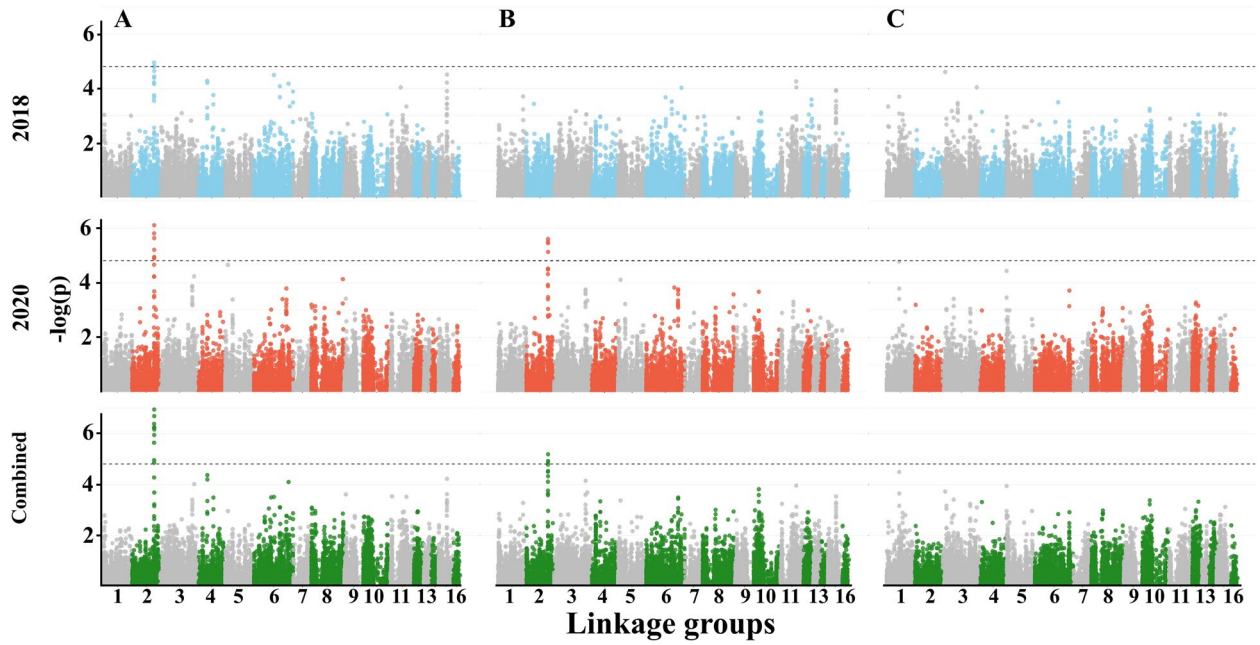


**Fig. 6** Manhattan plots for yield components traits. (**A**) seed-yield per plant, (**B**) seed number per plant, and (**C**) thousand-seed weight. The dashed line represents the genome-wide significance threshold

phenotypic responses within haplotypes (per genotype) in FD and SYPP across years. For FD, we found that genotypes in both haplotypes had a similar pattern ($P=0.5$ and $P=0.12$, for year effect) as expressed in parallel trend lines between genotypes across years (Fig. 7A). For SYPP, the differences in mean values across years were on the edge of statistical significance ($P=0.055$ and $P=0.32$ for Hap1 and Hap2, respectively), but the genotypes within each haplotype had different values across years as expressed with the crossing lines that connected the same genotypes between the two growing seasons (Fig. 7B).

To find the genetic basis underlying the biological influence on the traits, we searched for candidate genes (CGs) within this LG2 hotspot genomic region. Overall, we found 20 CGs: 7 CGs were located inside of the genomic region (i.e., within the significant SNPs boundaries),

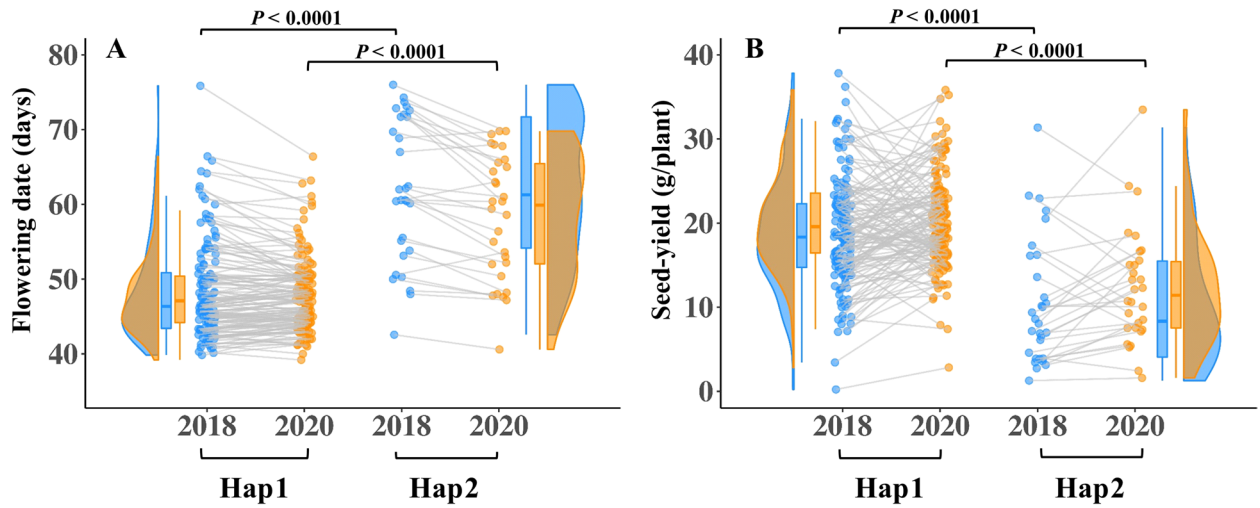


**Fig. 7** Haplotype analysis of the major genomic region on LG2. Analysis of variance of haplotypes allele on (**A**) flowering date and (**B**) seed-yield across two years. Every dot represents a genotype. Blue color represents the 2018 growing season, orange represents the 2020 growing season, and grey lines connect the same genotypes in the two growing seasons

while 13 CGs were consisted of upstream (8) and downstream (5) of the genomic region (Supplemental Table S10). All the 10 SNPs were located in regulatory or coding sequences of the CGs. Although four of the CGs encode uncharacterized proteins, most of the CGs were involved in signaling pathways and regulation and four of them (LOC105156148, LOC105156152, LOC105156159, and LOC105156313) were associated with controlling flowering time, floral development, and productivity in *Arabidopsis* and rice [46–49].

## Discussion

Understanding the genetic basis of agronomic traits is important for research and breeding efforts. Here, we explored the genetic and phenotypic variation in a newly established sesame panel (SCHUJI) aiming at dissecting the genomic architecture of morpho-agronomic traits. We hypothesized that as 'orphan crop-plant' sesame has not been subjected to modern selection, the geographically distributed panel preserved an ample allelic repertoire for agronomically important traits. Evaluation of the level and extent of genetic diversity among the sesame panel surprisingly showed that there is no clear geographical separation between genotypes (Fig. 3A). Similarly, previous genetic characterization of other sesame panels showed no clustering associated with continents and suggested sub-geographical regions or latitudes as key factors [19, 50]. Moreover, the ADMIXTURE analysis resulted in seven sub-populations (K = 7) with low to moderate differences between them (Fig. 3B; Supplemental Table S6). Thus, the lack of geographic signature in our, as well as other genetic panels, may be a consequence of recent genetic materials exchange [51] or unavailable information at the collection sites. As an example, even the smallest sub-populations (2, 3, and 4) that were more genetically uniform showed the various origin of genotypes (i.e., sub-population 4 included genotypes from Thailand, Israel, or unknown) (Fig. 3B).

### Wide phenotypic diversity among the sesame panel highlights the tradeoff between flowering date and productivity traits

High phenotypic diversity was found for all morpho-agronomical traits across the 2 years, with 2020 expressing higher values (Fig. 1). This observation agrees with previous studies using other sesame panels [19, 52]. In general, FD was found to be negatively correlated with yield components, such as RI, SYPP, and SNPP. On the other hand, late FD was associated with high NBPP, HTFC, and PH (Figs. 2, 4; Supplemental Table S4). Additional support for the key effect of FD on productivity comes from the k-means clustering analysis. This analysis clustered the panel into three groups (Fig. 2B), which were also associated with the genetic structure of the panel (Supplemental Fig. S4). Comparison between the mid-FD group (green) and the early FD (blue) showed that NBPP and PH have a positive effect on productivity traits, which is in line with previous reports in sesame [53]. Gadri et al. [7] showed in sesame that increasing the source size (i.e., vegetative organs) can support a higher seed set and filling. Thus, the higher yield potential of cluster 1 can be a consequence of greater vegetative biomass accumulation following more reproductive branches. Interestingly, the late-flowering group (red) had a similar number of branches as the intermediate group (green); however, most of them were not fertile due to the late-flowering phenotype. RI is a key trait for yield potential assessment and represents the ratio between RZ and PH. As expected, RI and FD are negatively correlated (Fig. 2). However, it is worth noted that a high value of RI can be a consequence of either high PH with low HTFC or small plants. Therefore, to promote higher yield potential, it is important to combine high RI with high RZ values.

Langham [5] reported that photoperiod responsiveness plays a key role in sesame flowering date and vegetative biomass accumulation. Our results show that while most genotypes that belong to the late FD and low RI group clustered together, they differ in their geographical origins (Fig. 3A; Supplemental Fig. S4), which may suggest the involvement of other genetic and/or environmental factors.

### Genetic architecture of agronomical traits reveals hotspot of overlapping genomic regions

To detect the genetic basis underlying observed phenotypic variation, we conducted single-marker regression GWAS. Overall, we detected 50 associated SNPs for all the traits, with 11 for 2018 and 19 for 2020 seasons, and 20 for the combined data that were spread along the entire sesame genome. While all the traits exhibited phenotypic variation (Fig. 1), we detected a relatively small number of SNPs associated with these traits, which could be the outcome of small population size ($n = 184$) or low genetic variation in QTL sites.

The flowering date was mapped to two genomic regions (13 SNPs) on LGs 2 and 11 that were consistent across years (Fig. 5A and Supplemental Table S9). The major genomic region on LG2 detected in the current study under the Mediterranean climate was previously reported in another sesame panel, however, it was less significant under other environments [19]. The advantage of using a diversity panel to detect new alleles is exemplified by comparing our results with a bi-parental population that was grown under the same environmental conditions. While Teboul et al. [3] detected six QTLs

using the $F_2$ population (S-91 × S-297), only the genomic region on LG11 was overlap with the current study.

The plant architecture traits were associated with 9 SNPs [HTFC (4), RZ (2), and RI (3)], with some overlaps (Fig. 5B, C; Supplemental Fig. S4A; Supplemental Table S9). Two genomic regions on LG2 and 16 showed overlaps for HTFC and RI, which correspond to a high genomic correlation ($r = -0.96$) between those traits. These results suggest that genotypes with shorter HTFC will have the potential to extend the growth period and develop more capsules (i.e., high RI). In addition, while HTFC and RZ had different mean values across years, the mean values of RI were similar, which pinpoint the potential of this trait in breeding programs to ensure high productivity. In the current study, we did not detect any genomic region associated with PH. This is a complex trait assembled through HTFC, RZ, and other morphological traits, such as internode number and length, that were not measured in this study. A previous study in sesame reported several QTLs, but most of them explained a small proportion of the phenotypic variation [54, 55], suggesting the complex nature of this trait (PH). While FD and PH showed positive phenotypic and genomic correlations (0.70 and 0.58, respectively), clustering analysis suggested that different genotypes with similar PH can differ in FD. Thus, the lack of genomic region associated with PH might result from many small effect loci or be masked by phenology. It is also worth noting that since sesame is grown as a summer crop under irrigation and characterized with indeterminate growth habits, various genetic, environmental, and management (i.e., G × E × M interactions) might affect the diversity of PH (Supplemental Table S3).

Twenty-eight SNPs were significantly associated with yield components. Of those, 20 were associated with SYPP, 7 for SNPP, and 1 for TSW. Seed size (TSW) is known to have moderate-high heritability in sesame [56, 57], as was found in the current study (0.88) and other crops, such as wheat [58] and pea (*Pisum sativum* L. [59]), was associated with only one region in 2020, with no overlap with SNPP and SYPP. Likewise, a small number of associated loci for TSW were found in bi-parental sesame populations [3, 60], which may suggest that this trait is under less complex genetic control or under the regulation of many small-effect genomic regions. The absence of significant genomic signals for TSW (as well as other traits in the current study) indicates that other approaches, such as a multi-locus GWAS model [61] or genomic prediction [62], may contribute to the understanding of the genetic basis. The positive phenotypic and genomic correlations between TSW and SYPP (0.37 and 0.44, respectively), and between SNPP and SYPP (0.86 and 0.93) on one side, and the absence of both correlations between TSW and SNPP (−0.04 and 0.17, respectively) on the other side, open up the possibility to breed simultaneously for both traits and improve yield.

SYPP and SNPP are highly polygenic traits associated with various anatomical and morphological traits (e.g., NBPP, RZ, RI, PH, number of capsules per plant, number of capsules per leaf axil, carpel number per capsule, and seed size). For SNPP and SYPP, we found one shared genomic region on LG2 overlapping with FD (Supplemental Table S9), which is supported by the negative phenotypic and genomic correlations between these traits (Fig. 4) and demonstrates the important role of FD on productivity. The effect of flowering on seed yield can be also attributed to the indetermine growth habit of sesame, and the agronomic practices to stop the irrigation to harvest the crop before it rains in autumn. Under such conditions, plants that were able to flower earlier had a longer period to produce flowers and capsules and obtain a higher yield. In contrast, genotypes that flowered later in the season were more exposed to environmental and management conditions. A similar pattern was found in other indetermined crops such as soybean [63] and chickpea (*Cicer arietinum* L. [64]). It is worth noted that while FD had high broad-sense heritability (0.97), both SYPP and SNPP had relatively lower broad-sense heritability estimates (0.74 and 0.66, respectively) (Table 1). These results, along with haplotypes analysis (Fig. 7), suggest that these yield components are controlled by other less heritable factors not related to FD.

In sesame, allelic variation within flowering-related genes contributes to variation in flowering date [65, 66]. In the current study, we found two genomic regions (LGs 2 and 11) that were associated with FD across years. The significant marker on LG11 causes a non-synonymous mutation in the first exon of the gene LOC105173174, which encodes to AT-hook motif nuclear-localized protein 9. The members of this gene family are associated with the regulation of flowering dates in other plants species [67]. The co-localization of the major genomic region on LG2 for FD and yield components together may suggest that this region contains one major gene with a pleiotropic effect or cluster of several genes. Analysis of CGs within this genomic region highlighted 20 flowering and productivity-related genes (Supplemental Table S10). LOC105156148 is encoding nitrate transporter (NRT1) and was found in *Arabidopsis* to interact with two flowering regulators transcription factors, *CONSTANS* and *FLOWERING LOCUS C* (FLC) [49]. LOC105156159 is encoding abscisic acid receptor *PYR1-like*, a mutant allele of this gene is found to be associated with growth and productivity in rice [48]. Two of the significant markers within this QTL were within the regulatory region (I.e., introns) of the LOC105156152 gene that encodes

*CHROMATIN REMODELING 19*. This gene was found regulating floral organs development in *Arabidopsis* [47]. Additional gene (LOC105156157) is translated to IAA-alanine resistance protein 1 and contains three SNPs in both regulatory and coding sequences. This gene was characterized in *Arabidopsis* and associated with auxin conjugate and homeostasis [68], and several mutants in this locus were associated with late flowering [69]. Yet, additional study is needed to test if contradictory genotypes will exhibit different expression levels or DNA polymorphisms (insertion/deletion) that were not detected in the current study.

The functional annotation of these genes and the co-localization of them in the same genomic region demonstrated how the variation in FD and SYPP could be genetically controlled together. Further investigation is needed to study how this genomic region (and the genes inside it) interact with other identified genomic regions for a deeper understanding of the genetic basis and mechanisms underlying FD and SYPP variations in sesame.

## Conclusions

While sesame is still mostly grown under a traditional cropping system, future sesame breeding targets should focus on improving yield and adaptively to more diverse environments. Thus, elucidating the genetic architecture controlling phenological, morphological, and yield components traits will aid in understanding selection criteria and better genetic-based breeding. Here we established a new sesame panel (SCHUJI) and explored its genetic variation for morpho-agronomic traits under the Mediterranean climate conditions. We showed the benefits of using the globally distributed panel for discovering new alleles associated with these traits. A major genomic region on LG2 was found in association with flowering date and yield components, indicating the crucial role of phenology on sesame production. A better understanding of the genetic variability underlying flowering date in sesame will serve as a basis for improving sesame adaptively to new cropping systems. Moreover, it will enhance breeding efforts and enable turning this important crop from domestically grown to global production in intensive agriculture.

#### Abbreviations

BLUE: Best linear unbiased estimate; CGs: Candidate genes; FD: Flowering date; GWAS: Genome-wide association studies; HTFC: Height to the first capsule; NBPP: Number of branches per plants; PH: Plant height; RI: Reproductive index; RZ: Reproductive zone; SNP: Single nucleotide polymorphism; SNPP: Seeds number per plant; SYPP: Seed-yield per plant; TSW: Thousand-seed weight.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-021-03328-4.

**Additional file 1:** Supplementary Tables. **Table S1**. List of the sesame lines from the SCHUJI panel and their geographical origin. **Table S2**. Cross-validation error for K=1-10. **Table S3.** Mean, standard deviations (SD) and coefficient of variation (CV) of the measured traits. **Table S4.** Phenotypic Correlations and *P*-values among the measured traits. **Table S5.** Cluster means from the K-means clustering analysis. **Table S6.** Fixation index (FST) values between the subpopulation in the panel. **Table S7.** Genomic heritability estimates of the measured traits. **Table S8.** Genomic correlations among the measured traits. **Table S9.** Significant genetic markers (SNPs) that are associated with the traits in the experiment. **Table S10.** Candidate genes for flowering date and seed-yield traits.

**Additional file 2: Fig S1**. Density distribution of reproductive zone measured in 2018 (blue) and 2020 (red).

**Additional file 3: Fig S2.** Genome-wide linkage disequilibrium pattern in the SCHUJI panel. The red dashed line represents the non-linear trend and the green dashed line indicates the point where linkage disequilibrium drops to half of its initial value (0.22) at 58,774 base pairs.

**Additional file 4: Fig S3.** Manhattan plots for (**A**) reproductive zone, (**B**) plant height and (**C**) number of branches per plant. Dashed line represents the genome-wide significant threshold.

**Additional file 5: Fig S4.** Principal component analysis on genetic markers of the SCHUJI panel. Colors represent different clusters from the k-means cluster analysis. Red color represents cluster 3, green color represents cluster 1 and blue color represents cluster 2.

#### Acknowledgments

We highly appreciate the excellent technical support of Dr. Ruchama Hayouka, Dr. Naama Teboul, Yaron Gadri, and Peleg's lab members.

#### Authors' contributions

IS, GM, and ZP designed the experiment. IS conducted the field experiment. IS, GM, and ZP analyzed the data and wrote the manuscript. The authors read and approved the final manuscript.

#### Funding

This research was partly supported by the Chief Scientist of the Israel Ministry of Agriculture and Rural Development (837-0150-14), a Research Grant from BARD, the United States - Israel Binational Agricultural Research and Development Fund (No. IS-5400-21), and The Hebrew University of Jerusalem.

#### Availability of data and materials

The genomic information can be found at https://figshare.com/s/a061d548a97237b51a61
The phenotypic information can be found at https://figshare.com/s/94a222afca9423d0b1aa

## Declarations

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.y.

Received: 7 June 2021 Accepted: 8 November 2021
Published online: 22 November 2021

## References


1. Bedigian D. Systematics and evolution in Sesamum L. (Pedaliaceae), part 1: evidence regarding the origin of sesame and its closest relatives. Webbia. 2015:1–42.
2. Anilakumar KR, Pal A, Khanum F, Bawa AS. Nutritional, medicinal and industrial uses of sesame (*Sesamum indicum* L.) seeds - An overview. Agric Conspec Sci. 2010;75(4):159–68.
3. Teboul N, Gadri Y, Berkovich Z, Reifen R, Peleg Z. Genetic architecture underpinning yield components and seed mineral-nutrients in sesame. Genes. 2020;11(10):1221.
4. Mushtaq A, Hanif MA, Ayub MA, Bhatti IA, Jilani MI. Sesame. Medicinal plants of South Asia. Elsevier. 2020; p. 601–15.
5. Langham R. Phenology of sesame. New crop new uses. 2007; p.144-182.
6. M . Delgado and D . M . Yermanos. Yield components of sesame (*Sesamum indicum* L.) under different population densities. Econ. Bot. 2017;29(1):69-78.
7. Gadri Y, Eshed Williams L, Peleg Z. Tradeoffs between yield components promote crop stability in sesame. Plant Sci. 2020;295:110105.
8. Mei H, Liu Y, Du Z, Wu K, Cui C, Jiang X, et al. High-density genetic map construction and gene mapping of basal branching habit and flowers per leaf axil in sesame. Front Plant Sci. 2017;8:636.
9. Miao H, Li C, Duan Y, Wei L, Ju M, Zhang H. Identification of a *Sidwf1* gene controlling short internode length trait in the sesame dwarf mutant dw607. Theor Appl Genet. 2020;133(1):73–86.
10. Tibbs Cortes L, Zhang Z, Yu J. Status and prospects of genome-wide association studies in plants. Plant Genome. 2021;e20077.
11. Guo Z, Chen D, Alqudah AM, Röder MS, Ganal MW, Schnurbusch T. Genome-wide association analyses of 54 traits identified multiple loci for the determination of floret fertility in wheat. New Phytol. 2017;214(1):257–70.
12. Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet. 2012;45(1):43–50.
13. Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH, et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat Commun. 2011;2(1):467.
14. Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol J. 2015;13(2):211–21.
15. Asekova S, Oh E, Kulkarni KP, Siddique MI, Lee MH, Kim JI, et al. An integrated approach of QTL mapping and genome-wide association analysis identifies candidate genes for phytophthora blight resistance in sesame (*Sesamum indicum* L.). Front Plant Sci. 2021;12:604709.
16. Li D, Dossa K, Zhang Y, Wei X, Wang L, Zhang Y, et al. GWAS uncovers differential genetic bases for drought and salt tolerances in sesame at the germination stage. Genes. 2018;9(2):87.
17. Dossa K, Li D, Zhou R, Yu J, Wang L, Zhang Y, et al. The genetic basis of drought tolerance in the high oil crop *Sesamum indicum*. Plant Biotechnol J. 2019;17(9):1788–803.
18. He Q, Xu F, Min MH, Chu SH, Kim KW, Park YJ. Genome-wide association study of vitamin E using genotyping by sequencing in sesame (*Sesamum indicum*). Genes Genom. 2019;41(9):1085–93.
19. Wei X, Liu K, Zhang Y, Feng Q, Wang L, Zhao Y, et al. Genetic discovery for oil production and quality in sesame. Nat Commun. 2015;6:8609.
20. Zhou R, Dossa K, Li D, Yu J, You J, Wei X, et al. Genome-wide association studies of 39 seed yield-related traits in sesame (*Sesamum indicum* L.). Int J Mol Sci. 2018;19(9):2794.
21. Dossa K, Diouf D, Wang L, Wei X, Zhang Y, Niang M, et al. The emerging oilseed crop *Sesamum indicum* enters the "Omics" era. Front Plant Sci. 2017;8:1154.
22. Core Team R. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2020. https://www.R-project.org/
23. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67:1–48.
24. Schmidt P, Hartung J, Bennewitz J, Hans-Peter P. Heritability in plant breeding on a genotype-difference basis. Genetics. 2019;212(4):991–1008.
25. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Softw. 2017;82(13):1–26.
26. Kassambara A, Mundt F. factoextra: Extract and visualize the results of multivariate data analyses. 2020. https://cran.r-project.org/package=factoextra
27. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. 2016.
28. Wei T, Simko V. Visualization of a correlation matrix [R package corrplot version 0.88]. 2021. Available from: https://cran.r-project.org/package=corrplot
29. Doyle J, Doyle J. Genomic plant DNA preparation from fresh tissue-CTAB method. Phytochem Bull. 1987;19:11–5.
30. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6(5):e19379.
31. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23(19):2633–5.
32. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97.
33. 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. Am J Hum Genet. 2007;81(3):559–75.
34. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64.
35. 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.
36. Weir BS, Cockerham CC. Estimating F-statisticts for the analysis of population structure. Evolution. 1984;38(6):1358–70.
37. Hill WG, Weir BS. Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol. 1988;33(1):54–78.
38. Marroni F, Pinosio S, Zaina G, Fogolari F, Felice N, Cattonaro F, et al. Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (*CAD4*) gene. Tree Genet Genom. 2011;7(5):1011–23.
39. Pérez P, De Los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014;198(2):483–95.
40. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23.
41. Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975;31(2):423.
42. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 2011;4(3):250–5.
43. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38(2):203–8.
44. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity. 2005;95(3):221–7.
45. Allen M, Poggiali D, Whitaker K, Marshall TR, van Langen J, Kievit RA. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res. 2021;4:63.
46. Ito T, Nagata N, Yoshiba Y, Ohme-Takagi M, Ma H, Shinozaki K. *Arabidopsis MALE STERILITY1* encodes a PHD-type transcription factor and regulates pollen and tapetum development. Plant Cell. 2007;19:3549–62.
47. Wu M-F, Sang Y, Bezhani S, Yamaguchi N, Han S-K, Li Z, et al. SWI2/SNF2 chromatin remodeling ATPases overcome polycomb repression and control floral organ identity with the LEAFY and SEPALLATA3 transcription factors. Proc Natl Acad Sci U S A. 2011;109(9):3576–81.
48. Miao C, Xiao L, Hua K, Zou C, Zhao Y, Bressan RA, et al. Mutations in a subfamily of abscisic acid receptor genes promote rice growth and productivity. Proc Natl Acad Sci U S A. 2018;15(23):6058–63.
49. Teng Y, Liang Y, Wang M, Mai H, Ke L. Nitrate Transporter 1.1 is involved in regulating flowering time via transcriptional regulation of FLOWERING LOCUS C in *Arabidopsis thaliana*. Plant Sci. 2019;284:30–36.
50. Dossa K, Wei X, Zhang Y, Fonceka D, Yang W, Diouf D, et al. Analysis of genetic diversity and population structure of sesame accessions from Africa and Asia as major centers of its cultivation. Genes. 2016;7(4):14.
51. Basak M, Uzun B, Yol E. Genetic diversity and population structure of the Mediterranean sesame core collection with use of genome-wide SNPs developed by double digest RAD-Seq. PLoS One. 2019;14(10):e0e223757.
52. Furat S, Uzun B. The use of agro-morphological characters for the assessment of genetic diversity in sesame (*Sesamum indicum* L.). Plant Omics Journal. 2010;3(3):85–91.

53. Baydar H. Breeding for the improvement of the ideal plant type of sesame. Plant Breed. 2005;124(3):263–7.
54. Wu K, Liu H, Yang M, Tao Y, Ma H, Wu W, et al. High-density genetic map construction and QTLs analysis of grain yield-related traits in sesame (*Sesamum indicum* L.) based on RAD-Seq technology. BMC Plant Biol. 2014;14(1):274.
55. Wang L, Xia Q, Zhang Y, Zhu X, Zhu X, Li D, et al. Updated sesame genome assembly and fine mapping of plant height and seed coat color QTLs using a new high-density genetic map. BMC Genomics. 2016;17(1):31.
56. Uzun B, Yol E, Furat S. Genetic advance, heritability and inheritance in determinate growth habit of sesame. Aust J Crop Sci. 2013;7:978–83.
57. Kalaiyarasi R, Lokeshkumar K, Mohanraj M, Priyadharshini A, Rajasekar R. Genetic variability parameters for yield and yield related traits in sesame (*Sesamum indicum* L.). Int J Curr Microbiol Appl Sci. 2019;8(8):819–25.
58. Sukumaran S, Lopes M, Dreisigacker S, Reynolds M. Genetic analysis of multi-environmental spring wheat trials identifies genomic regions for locus-specific trade-offs for grain weight and grain number. Theor Appl Genet. 2018;131(4):985–98.
59. Huang S, Gali KK, Tar'an B, Warkentin TD, Bueckert RA. Pea phenology: crop potential in a warming environment. Crop Sci. 2017;57(3):1540–51.
60. Du H, Zhang H, Wei L, Li C, Duan Y, Wang H. A high-density genetic map constructed using specific length amplified fragment (SLAF) sequencing and QTL mapping of seed-related traits in sesame (*Sesamum indicum* L.). BMC Plant Biol. 2019;19(1):588.
61. Berhe M, Dossa K, You J, Mboup PA, Diallo IN, Diouf D, et al. Genome-wide association study and its applications in the non-model crop *Sesamum indicum*. BMC Plant Biol. 2021;21(1):283.
62. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D. de los Campos G, et al. genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22:961–75.
63. Zhang J, Song Q, Cregan PB, Nelson RL, Wang X, Wu J, et al. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. BMC Genomics. 2015;16(1):217.
64. Upadhyaya HD, Bajaj D, Das S, Saxena MS, Badoni S, Kumar V, et al. A genome-scale integrated approach aids in genetic dissection of complex flowering time trait in chickpea. Plant Mol Biol. 2015;89(4–5):403–20.
65. Wei X, Zhu X, Yu J, Wang L, Zhang Y, Li D, et al. Identification of sesame genomic variations from genome comparison of landrace and variety. Front Plant Sci. 2016;7:1169.
66. Zhou R, Liu P, Li D, Zhang X, Wei X. Photoperiod response-related gene SiCOL1 contributes to flowering in sesame. BMC Plant Biol. 2018;18:343.
67. Zhao J, Favero DS, Qiu J, Roalson EH, Neff MM. Insights into the evolution and diversification of the AT-hook motif nuclear localized gene family in land plants. BMC Plant Biol. 2014;14(1):266.
68. Lasswell J, Rogg LE, Nelson DC, Rongey C, Bartel B. Cloning and characterization of IAR1, a gene required for auxin conjugate sensitivity in Arabidopsis. Plant Cell. 2000;12(12):2395–408.
69. Nelson DC, Lasswell J, Rogg LE, Cohen MA, Bartel B. FKF1, a clock-controlled gene that regulates the transition to flowering in *Arabidopsis*. Cell. 2000;101(3):331–40.


## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

**Learn more** biomedcentral.com/submissions