Song *et al. Horticulture Advances* (2025) 3:23
https://doi.org/10.1007/s44281-025-00078-3

Horticulture Advances

RESEARCH ARTICLE Open Access

Genomic analyses of arugula (*Eruca sativa*) accessions reveal domestication history and breeding insights

Shuang Song[1], Adrian M. J. Lee[1,2], Edwin J. K. Ong[1,2], Alden Y. Tan[1,2], Yi Lin Lim[1,2], Ting Xiang Neik[1,2], Seam Choon Law[1,2], Wan Zu Tang[1,2] and Fook Tim Chew[1,2,3*]

Abstract

Eruca sativa, or arugula, a leafy vegetable from the Mediterranean region, is gaining popularity worldwide. However, its domestication history and genetic diversity remain poorly understood. This study investigates the genetic background of globally collected arugula accessions, their domestication history, and the implications for future breeding strategies. Among the 202 arugula accessions analysed, four distinct groups were identified. Group 1 (G1) appears to be misidentified as arugula and is more closely related to radish (*Raphanus sativus*) and *Brassica* spp. Group 2 (G2), Group 3 (G3), and Group 4 (G4) likely represent arugula populations evolved separately, with G4 showing greater genetic isolation and drifts. Geographically, G2 accessions cluster in Europe, G3 in North Africa, and G4 in South Asia (furthest from the Mediterranean region). Phylogenetic and demographic history analyses suggest that G2 is more ancient than G3 and G4. While G3 and G4 share a more recent common ancestor, small scale introgression between G2 and G3 was detected. Selective sweeps signatures revealed 12 genes under positive selection in G4 compared to G2, and these 12 genes are likely associated with adaptive divergency, suggesting that arugula has not undergone stringent artificial selection. A potential hybrid zone between Turkey, Israel, and Iran was identified. Accessions from this zone may serve as a valuable resource for screening new arugula varieties with superior traits. Although 202 accessions were studied, some exhibited high genetic similarity. To enhance germplasm diversity for breeding, future efforts should prioritize expanding collections in regions where arugula has naturalized.

Keywords Genetic diversity, Domestication history, Hybrid zone, Adaptive divergence, Crop breeding

Introduction

Arugula (*Eruca sativa*) is a widely consumed leafy vegetable native to the Mediterranean region (Pignone 2011). It has also been naturalized in various parts of the world, including South Asia, where it is primarily cultivated for the production of the industrially important Taramira oil (Sastry 2003). The consumption of arugula dates back to ancient times, with records in the Bible and the Talmud (Yaniv 1996). In recent years, arugula is gaining popularity worldwide because it contains various health-promoting bioactive compounds, such as glucosinolates, flavonoids, and carotenoids, which contribute to its anti-carcinogenic and antimicrobial properties (Bell and Wagstaff 2019).

Although arugula has been cultivated since ancient times, its taxonomic classification remains ambiguous. *Eruca sativa* has also been referred to as *E. vesicaria* subsp. *sativa,* based on the assumption that *Eruca*

*Correspondence:
Fook Tim Chew
dbscft@nus.edu.sg
[1] Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543, Singapore
[2] NUS Agritech Centre (NAC) and Research NUS Agritech Centre (NAC) and Research Centre on Sustainable Urban Farming (SUrF), National University of Singapore, 85 Science Park Dr, #01-03, Singapore 118258, Singapore
[3] Lee Hiok Kwee Functional Genomics Laboratories, Block S2, Level 5, 14 Science Drive 4, Lower Kent Ridge Road, Singapore 117543, Singapore

© The Author(s) 2025. **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/.

is a monospecific genus (Hall et al. 2012). Pignone and Gómez-Campo (2011) proposed that the crop comprises three subspecies: subsp. *sativa*, subsp. *vesicaria*, and subsp. *pinnatifida*. Subsp. *vesicaria*, and subsp. *pinnatifida* are endemic to Spain and north-western Africa respectively (Warwick, et al. 2007), and have not been domesticated (Pignone and Gómez-Campo 2011). In contrast, (Greuter, et al. 2008) argued that there should be five species under *Eruca*, namely *E. loncholoma*, *E. pinnatifida*, *E. sativa*, *E. setulosa*, and *E. vesicaria*. Because of this taxonomic uncertainty, comprehensive genomic and phylogenetic survey has been recommended to clarify the diversity and relationships within *E. sativa/vesicaria* (Bell and Wagstaff 2019).

Despite its popularity and growing interest in cultivar development, artificial selection and breeding of arugula remain challenging. First, arugula is predominantly an outcrossing species with varying degrees of self-incompatibility (Pignone and Gómez-Campo 2011). Second, it is a diploid plant with 11 pairs of chromosomes (2n = 22) (Nothnagel et al. 2012), which is different from the closely related *Brassic*a spp. (AA, 2n = 16; BB, 2n = 18; and CC, 2n = 20) (Agrawal et al. 2020). Hybridization between arugula and *Brassica oleracea* has been achieved through protoplast fusion rather than conventional crossbreeding, which is commonly practiced for other species (Matsuzawa et al. 1999).

To better understand the domestication and diversity of arugula, we re-sequenced 202 globally collected arugula accessions and identified approximately 2 million high quality single nucleotide polymorphisms (SNPs). Although genetic makers have previously been developed to assess genetic diversity of arugula (Golkar and Bakhtiari 2020), the domestication history and breeding potential of its germplasm remains underexploited (Guijarro-Real et al. 2020). Through admixture analysis, gene flow analyses, and selective sweep scanning, our study provides new insights in the domestication and future breeding of arugula.

Materials and methods

Re-sequencing, reading mapping and SNP calling

A panel of 202 *Eruca sativa* accessions was obtained from various seed banks and commercial sources. One individual per accession was included in the re-sequencing process, with the exception of accession TUA13, for which one additional individual was also re-sequenced and included in downstream analyses to intermittently assess the reliability of the results. Leaf tissues were collected 28 days after sowing from each sample for DNA extraction. Library preparation followed Novogene's standard whole-genome sequencing protocol, and sequencing was performed on the Illumina NovaSeq™ 6000 platform. Approximately 4 GB of cleaned data were generated per sample, with a mean coverage depth of 7×. The cleaned reads were aligned to an in-house *Eruca sativa* reference genome (accession number: PRJNA1269263) using the Burrow-Wheeler Aligner (Li and Durbin 2009). Polymorphisms across all samples were called simultaneously using bcftools (Danecek et al. 2021) with the command "mpileup -Ou -q 20 -Q 30". SNP filtering was performed using PLINK (Chang et al. 2015), retaining only SNPs with a minor allele frequency (MAF) > 5% and a missing rate ≤ 0.02 per locus across all individuals. A more stringent missing rate filter (0.02 instead of 0.10) was applied to ensure that at least one individual from each admixture group was represented at each SNP position. Principal component analysis (PCA) analyses confirmed that using this more stringent missing rate filter did not substantially alter the population structure (Fig. S1).

Population structure and phylogenetic analyses

The population structure was first examined using ADMIXTURE (Alexander et al. 2009) with K values from 2 to 8. The lowest cross-validation error was at K = 4 (Fig. S2), suggesting the presence of four groups. To better illustrate the relationships among these groups, admixture analysis was rerun using 171 individuals with a dominant admixture group (> 0.6) following the approach of Kang et al. 2021). Visualization of the admixture results was performed using pong (Behr et al. 2016). PCA were performed in PLINK after linkage disequilibrium (LD) pruning ($r^2 > 0.2$). To visualise the demographic history of the arugula groups, Stairway Plot v2.0 was used based on joint site frequency spectrum (SFS) (Liu and Fu 2020). A generation time of 1 year was estimated since arugula is an annual crop. A mutation rate of 9.1×10^{-9} per site per year was estimated based on that of *Brassica rapa* (Park et al. 2019). The median and confidence intervals of the effective population size (N_e) were estimated from 200 bootstrap replicates. Pairwise F_{ST} values among the four groups, as well as the unassigned group, were calculated using VCFtools (Danecek et al. 2011). Rare allele frequency was calculated manually in Python, and inbreeding coefficients were computed using PLINK. The identity-by-state (IBS) kinship matrix between individuals was conducted using King (Manichaikul et al. 2010).

Phylogenetic tree analysis was performed using IQ-TREE (Nguyen et al. 2015). Ultrafast bootstrapping was performed with 1000 replicates, based on the best-fitting model (SYM + R6) as determined by the Akaike Information Criterion. The analysis also included known relatives of arugula, specifically *Aethionema arabicum*, *Eutrema salsugineum*, *Kernera saxatilis*, *Raphanus sativus*, *Erucastrum elatum*, *Brassica incana*, *Brassica villosa*, *B. oleracea*, *Brassica montana*, *Brassica*

cretica, *Brassica macrocarpa*, *Brassica rupestris*, *Brassica rapa*, and *E. sativa* published previously. *Aethionem arabicum* was used as the outgroup to root the tree. Additionally, one extra individual from *E. salsugineum* and *E. elatum* was included to assess whether they would be positioned adjacent to each other in the tree. The phylogenetic tree was visualized using Evolview (Subramanian et al. 2019).

Gene flow and migration events between G2, G3, and G4 were inferred using TreeMix (Pickrell and Pritchard 2012) without specifying a root group. The analysis was performed with the number of migration events varying from 0 to 10 (Fig. S3a). The optimal number of migration events was determined based on residual plots (Fig. S3b), selecting the model with minimal standard error and overfitting. Introgression among G2, G3, and G4 was inferred using Dsuite (Malinsky et al. 2021), with G1 as the outgroup. The same analysis was repeated for all four groups using the *Brassica* group as the outgroup.

Identifying candidate genes based on selective sweeps

Three methods were used to conduct selective sweep analysis between G2 and G4, namely F_{st} (Danecek et al. 2011), π ratio (G2/G4) (Danecek et al. 2011), and XP-CLR (Chen et al. 2010). Selective-sweep analyses were performed using sliding windows of 20 kb and a linkage disequilibrium threshold of 0.05. Selective sweeps were defined as the top 1% values for each analysis. The candidate genes were identified based on the overlapping signatures from all three analyses. The candidate genes were then queried against protein sequence databases using *Arabidopsis thaliana* homologs from GenBank to predict their functions. The string database was used to predict protein–protein interactions between the candidate genes and to detect potential Gene Ontology enrichment (Szklarczyk et al. 2019).

Unless specified, all visualization was performed in the R statistical computing environment (R Core Team 2023), using the RStudio integrated development environment (RStudio Team 2020).

Results

Genetic structure and phylogenetic inference

Admixture analysis identified four groups (Fig. 1a): Group 1 (G1, $n=4$), Group 2 (G2, $n=80$), Group 3 (G3, $n=11$), and Group 4 (G4, $n=76$). At K = 2, G4 first separated from G2, showing clear divergence between the two groups. At K = 3, G1 formed a distinct group, implying they have a unique ancestry not well captured by G2 or G4 alone. At K = 4, G3 appeared, mainly affecting accessions in between G2 and G4, indicating intermediate ancestry or gene flow between G3 and G2 or between G3 and G4. These results suggest that G2 is likely the ancestral group and that G4 is a relatively distinct and homogenous group.

To understand the evolutionary relationships of the 202 arugula accessions and their genetic divergence, a maximum likelihood phylogenetic tree was constructed. Since arugula belongs to the Brassicaceae family, eight *Brassica* spp., seven individuals from other species within the Brassicaceae family, and one previously published *E. sativa* individual were also included in the phylogenetic analysis to better elucidate the relationships between the arugula accessions collected in this study and the other known relatives. *Aethionema arabicum* was selected as the outgroup to root the tree because it is a known ancestral lineage within Brassicaceae (Kiefer et al. 2019). The phylogenetic tree reveals that G1 is likely misidentified as arugula, because individuals from G1 cluster in a different clade compared to the other three groups and separated from them by *Brassica* spp. (Fig. 1b). Therefore, there are likely three clusters of true arugula, which was also reported in a similar study using 155 arugula accessions (Guijarro-Real et al. 2020). The relative positioning of the true arugula accessions (G2, G3, G4, and the unassigned accessions) among other known species in the Brassicaceae family is consistent with previous findings, which indicates that *E. sativa* is more closely related to *Brassica* spp. than *Raphanus sativus* (Wang et al. 2014). While the earlier findings were based on mitochondrial genome resequencing, this study utilizes whole genome resequencing. Among the three arugula groups, G2,

(See figure on next page.)

Fig. 1 Genetic structure and phylogenetic relationships between the 202 accessions studied. **a** Admixture grouping inferred using ADMIXTURE for 171 accessions with K values ranging from 2 to 4. The 171 accessions were selected because they have a dominant genetic component (> 60%). **b** Maximum-likelihood phylogenetic tree of the 202 accessions inferred using IQ-TREE. The tree is a consensus tree constructed with 1000 Ultrafast Bootstrap (UFBoot) replicates, using *Aethionema arabicum* (Aarabicum) as the outgroup. Branch support values, representing UFBoot support percentages, are denoted by coloured dots on the branches. The analysis was performed using the best-fit substitution model selected by ModelFinder. The leaves of the phylogenetic tree are coloured according to their admixture group. The branch at which G2, G3, and G4 diverged from the other individuals is marked by □. The branch at which G2 diverged from G3 and G4 is marked by ☆. The branch at which G3 and G4 diverged is marked by △. Esalsugineum, *Eutrema salsugineum*; Ksaxatilis, *Kernera saxatilis*; Rsativus, *Raphanus sativus*; Eelatum, *Erucastrum elatum*; Bincana, *Brassica incana*; Bvillosa, *Brassica villosa*; Boleracea, *Brassica oleracea*; Bmontana, *Brassica montana*; Bcretica, *Brassica cretica*; Bmacrocarpa, *Brassica macrocarpa*; Brupestris, *Brassica rupestris*; Brapa, *Brassica rapa*; Esativa, *Eruca sativa*

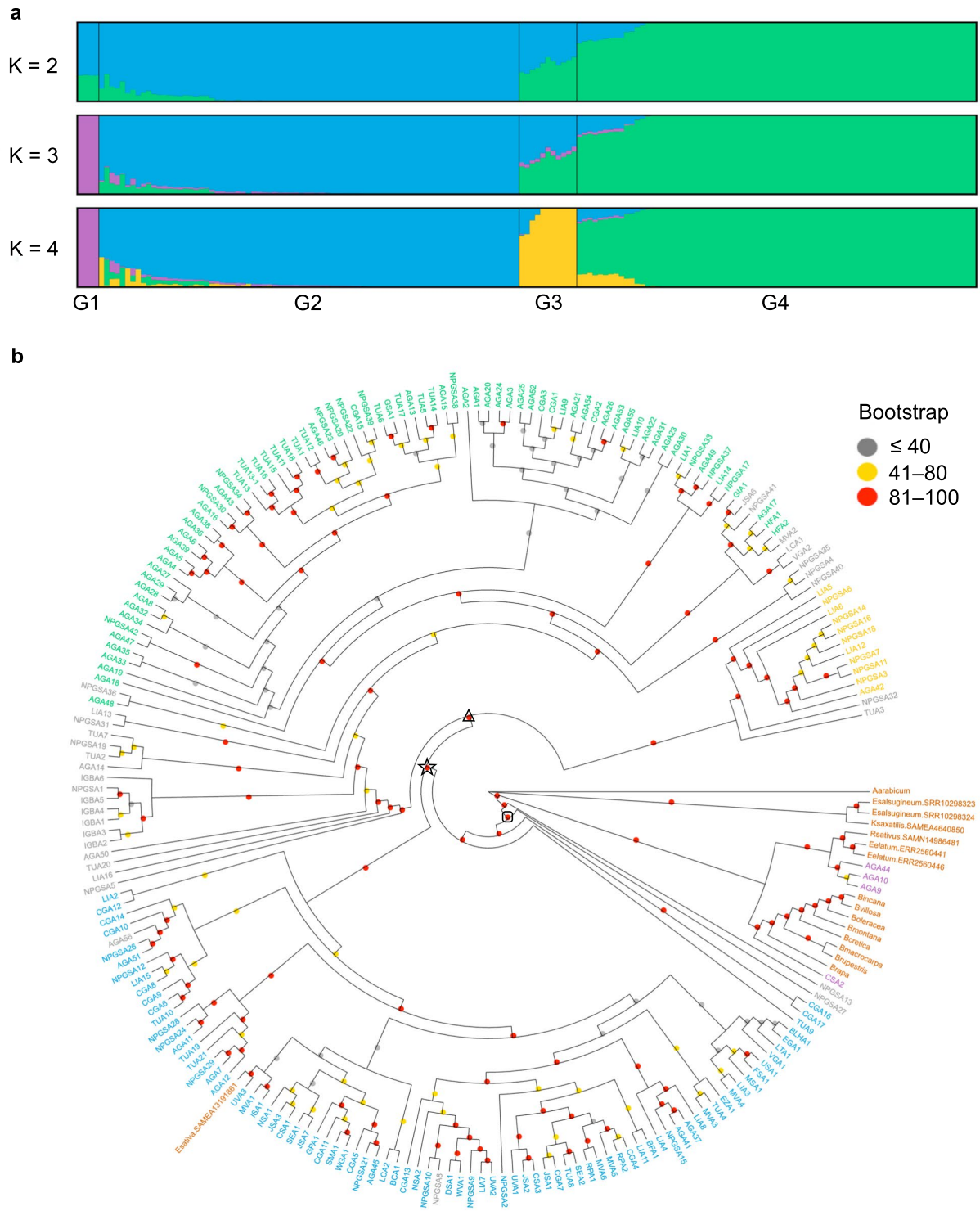


Fig. 1 (See legend on previous page.)

mainly clusters in Europe, is positioned closer to the root of the phylogenetic tree, suggesting that Europe may represent an early centre of cultivation, potentially marking the onset of domestication processes. This result is consistent with the ancient records of its cultivation in Rome and Greece (Renfrew and Sanderson 2012). However, since G3 and G4 are not positioned as subgroups of G2, substantial mutation or admixture events likely occurred, with new genetic materials introduced into G3 and G4 during the cultivation process. The accessions without an assigned group are likely hybrids because they are distributed in intermediate positions among the three groups. The clustering of the groups in the phylogenetic tree was also consistent with the PCA results, in which the G1 cluster was distinctly separated from the other three groups along PC2 (Fig. 2a), and in which accessions without an assigned group were found between the three main groups (Fig. 2b).

Geographical distribution, population diversity and genetic differentiation

Among the 202 accessions obtained, 159 had known origins. The remaining accessions without known origins are commercially purchased samples. The assigned admixture group and the country of origin for each accession can be found in Table S1. A total of 58 accessions were collected from Mediterranean-bordering regions, including Spain, Libya, Egypt, Israel, Syria, Cyprus, Turkey, Greece, Croatia, Italy, and France (Fig. 3a). Although we attempted to achieve global representation in arugula accessions, a large proportion (66 out of 202 accessions) originated from Pakistan and India in South Asia. These accessions were sourced from various gene banks, including the Australian Grains Genebank, the Leibniz Institute of Plant Genetic and Crop Plant Research, Tohoku University, the U.S. National Plant Germplasm System, and Wageningen University & Research. The

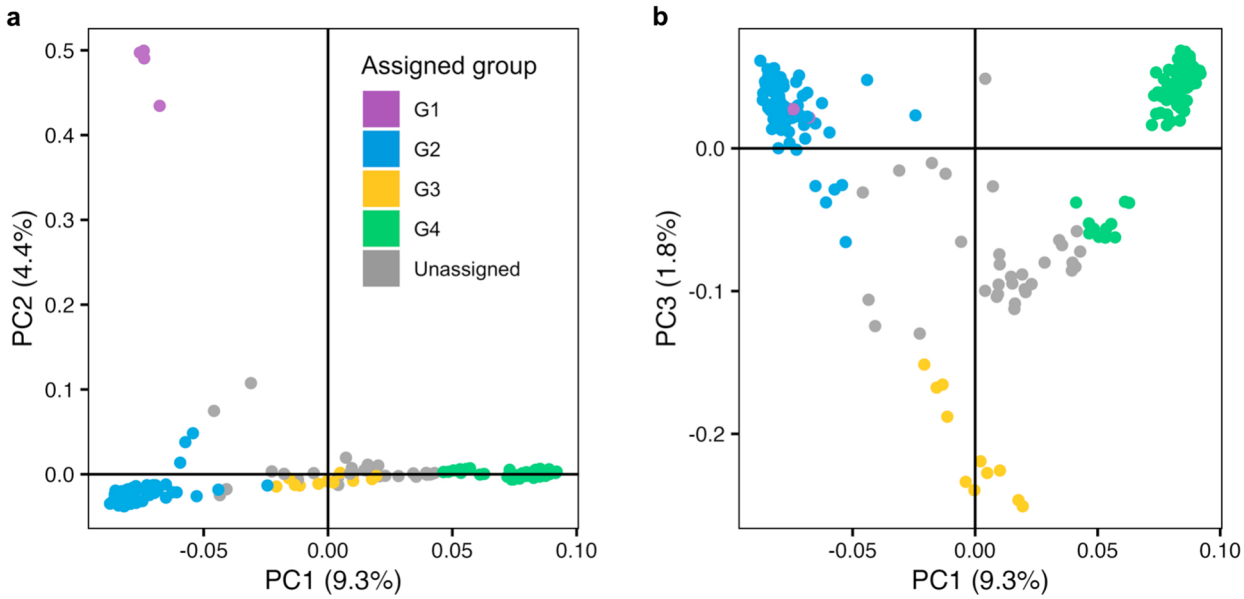


Fig. 2 Principal component analysis (PCA) plots showing the population structure among the four admixture groups. **a** Principal components (PCs) 1 and 2 showing the divergence and similarities among the admixture groups. **b** PCs 1 and 3 showing the divergence and similarities among the admixture groups

(See figure on next page.)

Fig. 3 Geographical distribution, population diversity, and demographic histories of the 202 arugula accessions studied. **a** Geographic distribution of the 159 arugula accessions with known origins. The regions were ordered according to their proximity to the Mediterranean Sea. Europe* excludes Spain. **b** Representative individuals from each admixture group with a scale bar (3.2 cm). For G1, accession AGA44 was used (more closely related to *Raphanus sativus*). **c** Demographic histories of the three arugula groups for the effective population size (Ne). g (generation time) = 1 years and μ (neutral mutation rate per generation) = 9.1×10^{-9}. **d** Rare allele analysis, population divergence represented by fixation index (F_{ST}), and inbreeding coefficient. The numbers inside the circles represent the % of rare alleles ($0.01 \leq MAF < 0.05$) among all single nucleotide polymorphisms (SNPs) within the subpopulation. The numbers next to the circle indicate the mean and standard deviation of the inbreeding coefficient for individuals within the subpopulation. The numbers on the arrows represent the pairwise F_{ST} values. **e** Identity-by-state (IBS) kinship matrix between individuals

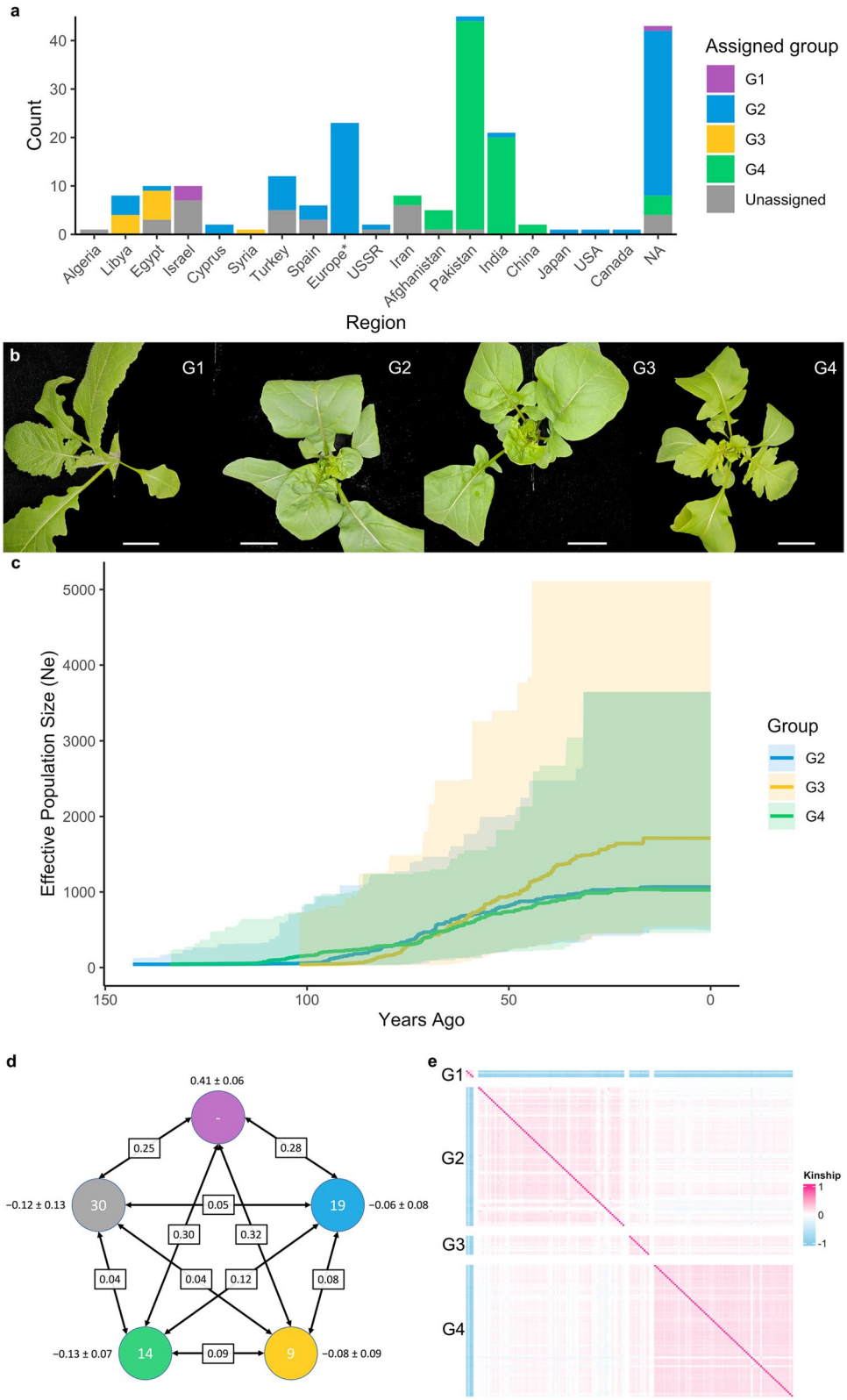


Fig. 3 (See legend on previous page.)

Mediterranean-bordering regions contained a mixture of G1, G2, and G3 groups. In contrast, G4 is limited to Asia, and most commercially purchased samples were assigned to G1, likely of European origin. In terms of the distribution of the four admixture groups, G1 is primarily found in Israel bordering eastern Mediterranean, G2 mainly distributed in Europe northwest of the Mediterranean, G3 predominantly located in North Africa along the southern Mediterranean coast, and G4 largely found in South Asia furthest from the Mediterranean region (Fig. 3a). Despite their geographic proximity, accessions without an assigned admixture group from Mediterranean regions were distributed across different positions in the phylogenetic tree. The Algerian accession (NPGSA13) was positioned between *Brassica* spp. and the arugula groups (Fig. 1b). NPGSA27 from Spain was also positioned between Brassica spp. and the arugula groups, but AGA50 and NPGSA5 were positioned between G2 and G4. Egyptian accessions NPGSA32 and TUA3 were positioned adjacent to the G3 cluster. The seven unclassified accessions from Israel and the five unclassified accessions from Turkey were placed between G2 and G4.

Morphologically, G1 leaves are hairier, with more serrated leaf edges and more prominent venation (Fig. 3b). In contrast, morphological differences among the other three groups at the vegetative stage were less pronounced. However, the three groups have been reported to be valued for different traits. G2 and G3 mainly use for leaf consumptions while G4 is used for seed oil production (Pignone and Gómez-Campo 2011; Warwick et al. 2007; Zafar-Pashanezhad et al. 2020).

Although we could not precisely date the divergence of the three arugula groups, the Stairway Plot analysis suggested that G2, primarily from Europe, is likely the most ancient lineage (Fig. 3c), which is consistent with the results of the phylogenetic tree (Fig. 1b). Additionally, all three groups showed evidence of past population expansion followed by stabilization, with G3 having higher effective population size (N_e) than G2 and G4. However, the timescale of the Stairway Plot may not be accurate because of the uncertainty in the estimated mutation rate.

Negative inbreeding coefficients were observed for G2, G3, and G4 (Fig. 3d), consistent with the fact that arugula is a cross-pollinated species commonly bred through mass selection of desirable traits (Morales and Janick 2002). Despite having a large number of accessions, G2 and G4 had higher percentages of rare alleles possibly because of their relatively small effective population sizes (Fig. 3c) or may have undergone more relaxed purifying selection. In contrast, G3 had bigger effective population size (Fig. 3c), leading to a lower percentage of rare alleles. Since G1 is not a true arugula group, it is the most genetically distinct from the other 3 admixture groups based on pairwise fixation index (F_{ST}) values (Fig. 3d) and likely representing a relatively isolated subpopulation with relatively high homozygosity (inbreeding coefficient: 0.41 ± 0.06). This high homozygosity is unusual for arugula because it is mostly self-incompatible (Pignone and Gómez-Campo 2011).

The (Identity-by-state) IBS kinship matrix showed that for all groups the within-group coefficients were substantially higher than the between-group coefficients (Fig. 3e). The within-group kinship coefficients of G4 were higher than the rest, showing that the individuals in G4 are more closely related. G4 appeared to be more closely related to G3 than to G2, and similarly, G2 is also more closely related to G3 than G4, possibly indicating that G3 is a connective group that potentially retained ancestral variations shared by both G2 and G4.

Intraspecific gene flow

TreeMix gene flow analysis suggests that there was no historical allele migration among the three groups across zero to ten migration events (Fig. 4a and Fig. S3). Nonetheless, consistent with the phylogenetic tree (Fig. 1b), G3 and G4 share the most recent common ancestor. Introgression was detected between G2 and G3 (D-statistic = 0.06 and Z = 9.47) but not between G2 and G4 (Fig. 4b and Table S2). The absence of introgression between G2 and G4 may reflect differing breeding priorities or major geological barriers, despite the presence of potential hybrids between these two groups (*i.e.*, unassigned accessions positioned intermediately between the two groups in Fig. 1b). The gene flow between G2 and G3 may be small scale since the gene flow was only detected by ABBA-BABA but not by Treemix (Fig. 4), suggesting limited interaction between the groups since their establishment in the different geological regions.

Genome-wide selective sweep signatures

Because G4 diverged more recently from the shared common ancestor than G2, and the two groups have comparable sample sizes, they were compared to identify genomic regions potentially associated with arugula domestication. Three approaches were used, namely weighted F_{st}, π ratio (G2/G4), and XP-CLR. The top 1% of the genomic windows shared among the three analyses were considered selective sweeps and genes overlapped with the genomic windows were extracted (Fig. 5). A total of 12 shared genes were identified across the three analyses. Protein–protein interaction analysis of these 12 genes revealed no significant GO enrichment (*P*-value = 0.68), although *RGGC* and *RPL24* showed potential co-expression (Fig. S4). These genes do not appear to include typical domestication genes, such as genes regulating flowering time (Xu et al. 2022) or

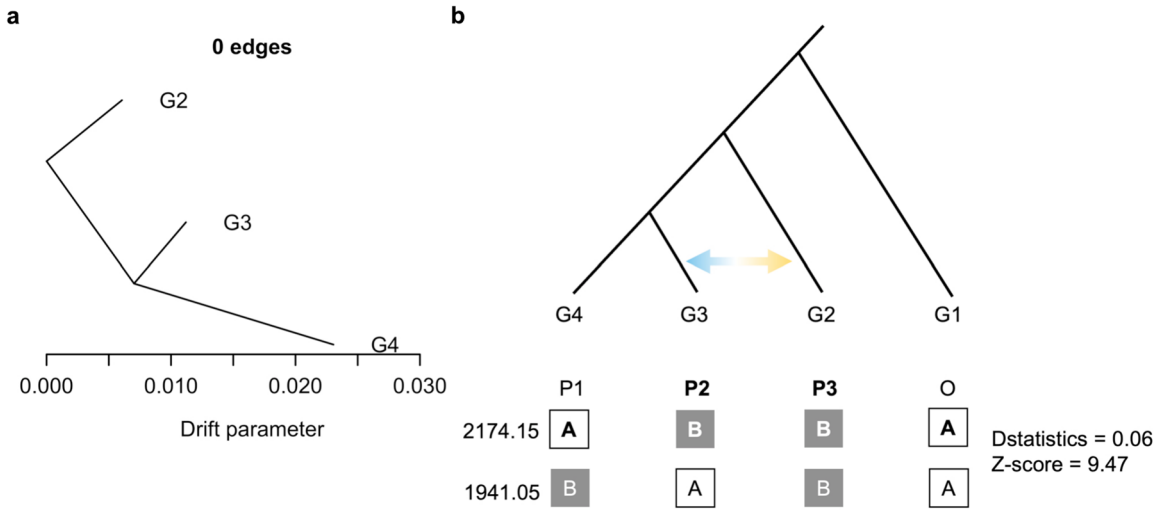


Fig. 4 Intraspecific gene flow. **a** Gene flow among the three arugula admixture groups studied. The population tree was inferred using TreeMix analysis. The tree is unrooted with zero edge. **b** Results of the ABBA-BABA (D-statistic) test, used to detect introgression among populations

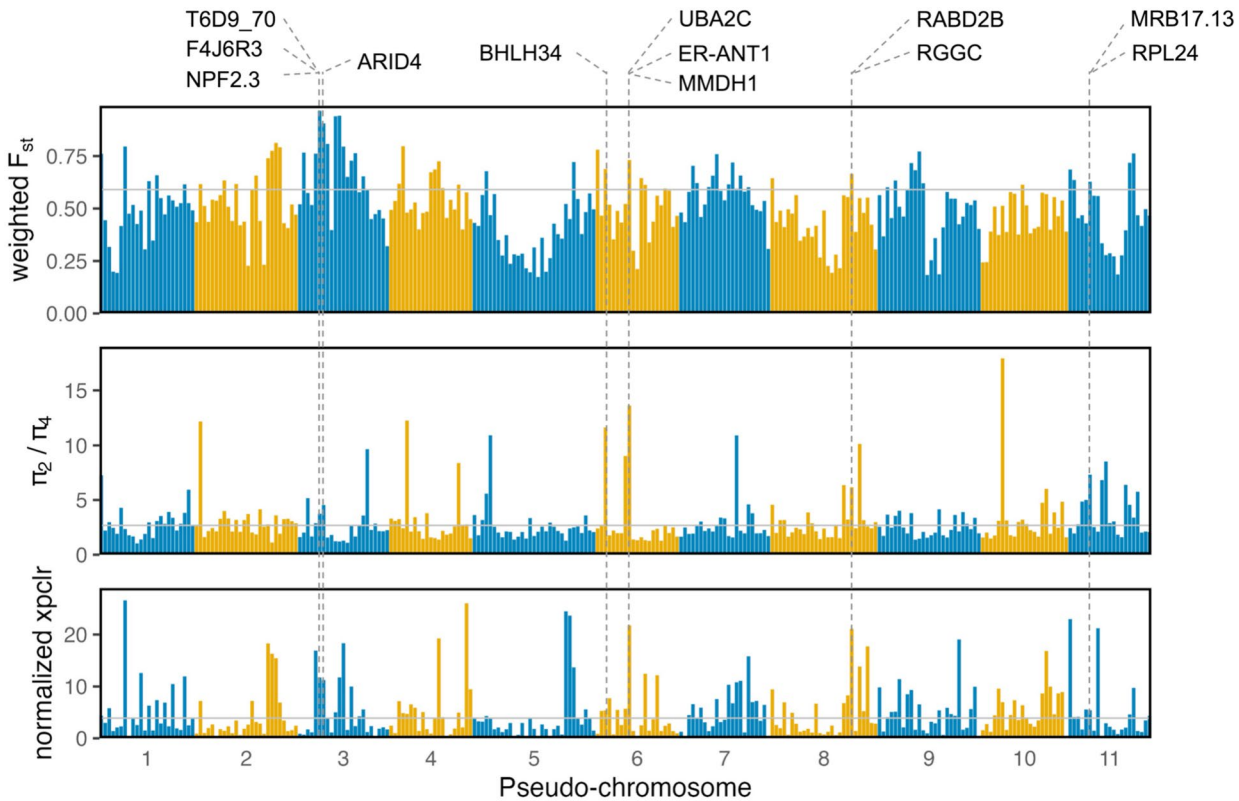


Fig. 5 Selective sweep scanning between G2 and G4 using weighted F_{st}, π ratio (π_2/π_4), and XP-CLR analyses. The horizontal grey line corresponds to the top 1% values for each analysis. The candidate genes were identified based on the overlapping signatures from all three analyses. ER-ANT1, Endoplasmic reticulum-adenine nucleotide transporter 1; MRB17.13, Zinc finger protein-like protein; BHLH34, Basic helix-loop-helix DNA-binding superfamily protein; MMDH1, Lactate/malate dehydrogenase family protein. UBA2C, RNA-binding (RRM/RBD/RNP motifs) family protein; T6D9_70, Hydrolase family protein/HAD-superfamily protein; F4J6R3, Nucleotidyl transferase family protein; NPF2.3, Protein NRT1/PTR FAMILY 2.3; RPL24, Plastid ribosomal protein L24; RGGC, RGG (arginine-glycine-glycine) repeats nuclear RNA binding protein C; RABD2B, Ras-related protein RABD2b; ARID4, AT-rich interactive domain-containing protein 4

seed shattering (Liu et al. 2018). Instead, two genes are potentially involved in defence responses (*ER-ANT1* and *UBA2C*), *BHLH34* in seed germination, *MMDH1* in photorespiration, *NPF2.3* in nitrate assimilation, and *RABD2B* in pollen tube growth (Table S3).

Discussion

Population structure and geographic isolation among the groups of *Eruca sativa*

Clear population structure and differentiation among the three admixture groups were observed with two major groups in Europe (G2) and South Asian (G4), respectively, and a smaller group in the South and Eastern Mediterranean regions (G3). A potential hybrid zone was identified between Turkey, Israel, and Iran, where accessions without an assigned admixture group were positioned between G2 and G4 or between G4 and G3 (Fig. 1b). The presence of this hybrid zone may have facilitated genetic exchange between G2 and G3, as detected by the ABBA-BABA analysis. However, no gene flow was detected between G4 and the other two groups, suggesting that G4 may be geographically isolated. This isolation likely stems from differences in cultivation purposes: while arugula is primarily consumed as a leafy vegetable in the Mediterranean and West Asia, it is mainly cultivated in South Asia, especially India and Pakistan, for seed oil production (Sihag 2023).

Potential spread routes of *Eruca sativa*

Despite the geographical proximity of China and Japan, the arugula accessions from China were assigned to G4, while those from Japan was assigned to G2 (Fig. 3a). This suggests two distinct introduction routes for arugula from the Mediterranean region. The eastward dispersal of arugula from the Mediterranean region via South Asia to China may be linked to Silk Road trade, a major trade route for trans-Eurasian crop exchange established since prehistory (202 BC to AD 220) (Dong et al. 2022). In contrast, the westward movement of arugula to North America and later Japan likely occurred after European colonization—first with Spanish arrival in the Americas in 1492 (Traboulay 1994) and later Portuguese arrival in Japan in 1543 (Lidin 2003). The eastward route is likely older, given that arugula was documented as a naturalized medicinal plant in Yunnan, southwestern China by 1476 (Lan 1959). However, the stairway plot (Fig. 3c) indicates that the divergence of the three arugula groups occurred relatively recently (~150 years ago). The discrepancy between historical records and genetic estimates could be caused by 1) misidentification, *i.e.*, the traditional medicine documented is not arugula (similar to the misidentified G1 admixture group) or 2) an inaccurate neutral mutation rate used in the stairway plot. Nevertheless, archaeological and genetic evidence to confirm these hypotheses is currently lacking.

Genetic adaptation and breeding implications

Selective sweep analysis suggests that arugula has potential for stringent artificial selection. Twelve candidate genes were identified as under selection, but they are more likely associated with adaptive divergence than direct artificial selection, indicating that the studied arugula accessions primarily represent ecotypes or landraces. This highlights their untapped potential for trait improvement through artificial breeding. Notably, G4 accessions are believed to have superior disease resistance, possibly due to the defence-related genes identified in the selective sweep analysis (Sastry 2003). These genes could be valuable for introgression into other arugula varieties or related *Brassica* crops to improve their disease resistance. However, artificial breeding of arugula is more challenging than in other *Brassica* species, as interspecific hybridization between arugula and related taxa is rare due to incompatibility mechanisms (Sun et al. 2005). To date, successful intergeneric hybrids between arugula and related *Brassica* species have only been achieved using embryo rescue or protoplast fusion (Olesen et al. 2012; Dai et al. 2004). This limitation suggests that intraspecific hybridization among the three arugula groups (G2, G3, and G4) may be a more feasible strategy to improve agronomic traits of arugula for different uses, such as leaf production or seed oil extraction. Accessions without an assigned admixture group (those not grouped into G1, G2, G3, and G4) may constitute an underutilized genetic resource for future arugula breeding, as they are likely natural hybrids among the groups and with wild relatives. For example, the two unassigned accessions (NPGSA13 and NPGSA27), which cluster between G2 and *Brassica* spp. (Fig. 1b), might be wild arugula relatives. This hypothesis is supported by their collection sites—Algeria and Spain—regions proposed as diversification centres for arugula (Pignone and Gómez-Campo 2011; Warwick et al. 2007; Hanelt et al. 2001). However, the Mediterranean region is underrepresented in this study, whereas South Asia is overrepresented—more than a quarter of the accessions (66 out of 202) were unintentionally collected from South Asia despite our attempt to achieve global representation. To address this, greater effort should be directed towards collecting genetically diverse accessions from regions where arugula has naturalized, thereby broadening the genetic base of available germplasm.

In this study, all accessions were obtained directly from various seed banks or commercial sources without additional screening, and genetically similar or identical accessions may have been unintentionally included

multiple times due to the following reasons. Firstly, seeds collected from nearby locations are often highly similar although they are labelled as different accessions; secondly, commercial seed suppliers may have obtained their seeds from seed banks; lastly, the seed banks exchange seeds, and the lack of a universal index number means the same accession in different seed banks could be misclassified as distinct. We included two individuals from accession TUA13 in multiple analyses to intermittently assess the reliability of the results. Although these two individuals typically clustered together or adjacent to each other in most analyses, the phylogenetic tree revealed that one TUA13 individual was more closely related to TUA16 (IBS kinship: 0.23) than to the other individual from TUA13 (IBS kinship: 0.21), suggesting a close relationship between the two accessions. While including genetically similar accessions could introduce minor bias in population structure inference and diversity estimates, this is unlikely to hinder future breeding efforts. The observed genetic clustering remains useful in guiding parental selection and germplasm management by identifying distinct family groups and breeding line.

Conclusions

This study evaluated the genetic diversity of global arugula (*Eruca sativa*) germplasm. Although native to the Mediterranean region, arugula is now widely cultivated in Europe (G2) for leaf consumption and South Asia (G4) for seed oil production. We identified three main genetic clusters with distinct genotypes and geographical distributions, likely reflecting independent evolutionary trajectories. However, a small-scale gene flow between G2 (Europe) and G3 (North Africa) was detected. Future studies could investigate the traits under selection in each group and the underlying genetic mechanisms controlling these traits. Our findings also suggest that the genetic differences between the three groups were mainly a result of adaptive divergence and arugula has yet to undergo stringent artificial selection, underscoring its potential for further trait improvement. Notably, admixed accessions (accessions without assigned admixture group) found in a potential hybrid zone between Turkey, Israel, and Iran — where hybridization between G2, G3 and G4 has occurred — represent a potential source of desirable traits because of the high genetic divergence and genetic diversity. These accessions warrant further attention as potential parental lines with novel trait combinations for future breeding programs. Despite the widespread cultivation of arugula in Europe and South Asia, the effective population sizes of the two groups remain relatively small. Greater effort should focus on collecting genetically diverse accessions from regions where arugula has naturalized, in order to broaden the genetic diversity of the germplasm.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s44281-025-00078-3.

Supplementary Material 1: Fig. S1. Population structure analysis of arugula accessions. a Population structure of the 202 arugula accessions, represented by PC1 and PC2 using PLINK filter (–geno = 0.02). b Population structure of the 202 arugula accessions, represented by PC1 and PC2 using a more relaxed PLINK filter (–geno = 0.10).

Supplementary Material 2: Fig. S2. Cross-validation (CV) errors for the admixture analysis of 203 individuals from 202 arugula accessions, with K values ranging from 2 to 8.

Supplementary Material 3: Fig. S3. Treemix results among G2, G3 and G4. a TreeMix plots showing migration events (0 to 10) among G2, G3, and G4. b Residual plots corresponding to a.

Supplementary Material 4: Table S1. Admixture grouping and country of origin for each accession included in this study.

Supplementary Material 5: Table S2. D-statistics for multiple triplet combinations among G1, G2, G3, and G4 with *Brassica* group being the outgroup.

Supplementary Material 6: Fig. S4. Visualization of the 12 identified genes using string database. Filled nodes indicate the protein's structure is known or predicted. Coloured nodes suggest first shell of interactors. Black line indicates co-expression and purple line indicates experimentally determined known interaction. ER-ANT1, Endoplasmic reticulum-adenine nucleotide transporter 1; MRB17.13, Zinc finger protein-like protein; BHLH34, Basic helix-loop-helix DNA-binding superfamily protein; MMDH1, Lactate/malate dehydrogenase family protein; UBA2C, RNA-binding (RRM/RBD/RNP motifs) family protein; T6D9_70, Hydrolase family protein/HAD-superfamily protein; F4J6R3, Nucleotidyltransferase family protein; NPF2.3, Protein NRT1/PTR FAMILY 2.3; RPL24, Plastid ribosomal protein L24; RGGC, RGG (arginine-glycine-glycine) repeats nuclear RNA binding protein C; RABD2B, Ras-related protein RABD2b; ARID4, AT-rich interactive domain-containing protein 4.

Supplementary Material 7: Table S3. Detailed information on the 12 potential genes identified to be positively selected in G4. The biological processes regulated by the genes were predicted by UniProt.

Acknowledgements

The authors would like to thank National University of Singapore (NUS) Enterprise for providing space and facilities at NUS Agritech Center (NAC) for the team to conduct growth experiments. F.T.C. received grants from National University of Singapore [N-154-000-038-001 (E-154-00-0017-01), C141-000-077-001 (E-141-00-0096-01)], the Singapore Ministry of Education Academic Research Fund (R154-000-191-112, R-154-000-404-112, R-154-000-553-112, R-154-000-565-112, R-154-000-630-112, R-154- 000-A08-592, R-154-000-A27-597, R-154-000-A91-592, R-154-000-A95-592, R154-000-B99-114), Biomedical Research Council (BMRC) (Singapore) (BMRC/01/1/21/18/077, BMRC/04/1/21/19/315, BMRC/APG2013/108), Singapore Immunology Network (SIgN-06-006, SIgN-08-020), National Medical Research Council (NMRC) (Singapore) [NMRC/1150/2008, OFIR-G20nov-0033, MOH-001636 (OFLCG23may-0038, A-8002641-00-00)], the National Research Foundation (NRF) (Singapore) (NRF-MP-2020-0004), Singapore Food Agency (SFA) (SFS_RND_SUFP_001_04, W22W3D0006), Singapore's Economic Development Board (EDB) (A-8002576-00-00), and the Agency for Science Technology and Research (A*STAR) (Singapore) (H17/01/a0/008, APG2013/108).

Authors' contributions

S.S., conceptualization, methodology, investigation, formal analysis, data curation, visualisation, writing— original draft, review & editing. A.M.J.L., E.J.K.O., A.Y.T. & Y.L.L., methodology, investigation, data curation, writing—review & editing. T.X.N., S.C.L. & W.Z.T., data curation, writing—review & editing. F.T.C., supervision, resources, project administration, funding acquisition, writing—review & editing. All authors have read and approved the final manuscript.

Funding
This project was funded by the Singapore Food Agency under grant number SFS_RND_SUFP_001_04.

Data availability
Data will be shared upon request by the readers.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
F.T.C. reports grants from the National University of Singapore, the Singapore Ministry of Education Academic Research Fund, Singapore Immunology Network, National Medical Research Council (NMRC) (Singapore), Biomedical Research Council (BMRC) (Singapore), the National Research Foundation (NRF) (Singapore), Singapore Food Agency (SFA), Singapore's Economic Development Board (EDB), and the Agency for Science Technology and Research (A*STAR) (Singapore), during the conduct of the study; and consulting fees from Sime Darby Technology Centre; First Resources Ltd; Genting Plantation, Olam International, Musim Mas, and Syngenta Crop Protection, outside the submitted work. The funding agencies had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The other authors declare no other competing interests.

Received: 7 March 2025 Revised: 12 June 2025 Accepted: 18 June 2025
Published online: 20 August 2025

References

Agrawal N, Gupta M, Banga SS, Heslop-Harrison JP. Identification of chromosomes and chromosome rearrangements in crop brassicas and *Raphanus sativus*: a cytogenetic toolkit using synthesized massive oligonucleotide libraries. Front Plant Sci. 2020;11: 598039. https://doi.org/10.3389/FPLS.2020.598039.

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64. https://doi.org/10.1101/gr.094052.109.

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics. 2016;32(18):2817–23. https://doi.org/10.1093/bioinformatics/btw327.

Bell L, Wagstaff C. Rocket science: a review of phytochemical & health-related research in *Eruca* & Diplotaxis species. Food Chemistry: X. 2019;1: 100002. https://doi.org/10.1016/j.fochx.2018.100002.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):7. https://doi.org/10.1186/s13742-015-0047-8.

Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Res. 2010;20(3):393–402. https://doi.org/10.1101/gr.100545.109.

Dai L, Li X, Guan C, Zhong J. On ovary and embryo culture of hybrid between rapeseeds and Rocketsalad. J Hunan Agric Univ (Nat Sci). 2004;3:201–4.

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. https://doi.org/10.1093/bioinformatics/btr330.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008. https://doi.org/10.1093/GIGASCIENCE/GIAB008.

Dong G, Du L, Yang L, Lu M, Qiu M, Li H, et al. Dispersal of crop-livestock and geographical-temporal variation of subsistence along the Steppe and Silk Roads across Eurasia in prehistory. Sci China Earth Sci. 2022;65(7):1187–210. https://doi.org/10.1007/s11430-021-9929-x.

Golkar P, Bakhtiari MA. Evaluation of genetic diversity in the world collection of *Eruca sativa* L. using oil content, fatty acids and molecular markers. Ind Crops Prod. 2020;148: 112280. https://doi.org/10.1016/j.indcrop.2020.112280.

Greuter W, Burdet HM, Long G. Med-Checklist: a critical inventory of vascular plants of the circum-mediterranean countries, vol. 3. Geneva: OPTIMA and Conservatoire et Jardin botaniques; 2008.

Guijarro-Real C, Navarro A, Esposito S, Festa G, Macellaro R, Di Cesare C, et al. Large scale phenotyping and molecular analysis in a germplasm collection of rocket salad (*Eruca vesicaria*) reveal a differentiation of the gene pool by geographical origin. Euphytica. 2020;216(3):53. https://doi.org/10.1007/s10681-020-02586-x.

Hall M, Jobling J, Rogers G. Factors affecting growth of perennial wall rocket and annual garden rocket. Int J Veg Sci. 2012;18(4):393–411. https://doi.org/10.1080/19315260.2012.660565.

Hanelt P, Buttner R, Mansfeld R. Mansfeld's encyclopedia of agricultural and horticultural crops (except ornamentals). Springer; 2001. https://doi.org/10.1007/978-3-540-30442-5.

Kang L, Qian L, Zheng M, Chen L, Chen H, Yang L, et al. Genomic insights into the origin, domestication and diversification of *Brassica juncea*. Nat Genet. 2021;53(9):1392–402. https://doi.org/10.1038/s41588-021-00922-y.

Kiefer C, Willing EM, Jiao WB, Sun H, Piednoel M, Humann U, et al. Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation. Nat Plants. 2019;5(8):846–55. https://doi.org/10.1038/s41477-019-0486-9.

Lan M. Dian Nan Ben Cao. 1959.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Lidin OG. Tanegashima-the arrival of Europe in Japan. Routledge; 2003.

Liu H, Li Q, Xing Y. Genes contributing to domestication of rice seed traits and its global expansion. Genes. 2018. https://doi.org/10.3390/genes9100489.

Liu X, Fu YX. Stairway plot 2: demographic history inference with folded SNP frequency spectra. Genome Biol. 2020;21(1):280. https://doi.org/10.1186/s13059-020-02196-9.

Malinsky M, Matschiner M, Svardal H. Dsuite-fast *D*-statistics and related admixture evidence from VCF files. Mol Ecol Resour. 2021;21(2):584–95. https://doi.org/10.1111/1755-0998.13265.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867–73. https://doi.org/10.1093/bioinformatics/btq559.

Matsuzawa Y, Mekiyanon S, Kaneko Y, Bang SW, Wakui K, Takahata Y. Male sterility in alloplasmic *Brassica rapa* L. carrying *Eruca sativa* cytoplasm. Plant Breed. 1999;118(1):82–4.

Morales M, Janick J. Arugula: a promising specialty leaf vegetable. In: Janick J, Whipkey A, editors. n.d.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74. https://doi.org/10.1093/molbev/msu300.

Nothnagel T, Budahn H, Schrader O, Klocke E, editors. Intergeneric hybridization between *Brassica oleracea* and *Eruca sativa*. VI International Symposium on Brassicas and XVIII Crucifer Genetics Workshop 1005; 2012.

Olesen MH, Romme HA, Kampmann HH. Rucola plants with cytoplasmic male sterility (CMS). US8247655B2. Google; 2012.

Park JS, Park JH, Park YD. Construction of pseudomolecule sequences of *Brassica rapa* ssp. *pekinensis* inbred line CT001 and analysis of spontaneous mutations derived via sexual propagation. PLoS One. 2019;14(9):e0222283. https://doi.org/10.1371/journal.pone.0222283.

Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8(11): e1002967. https://doi.org/10.1371/journal.pgen.1002967.

Pignone D, Gómez-Campo C. Eruca. Wild crop relatives: genomic and breeding resources. In: Kole C, editor. Springer; 2011. p. 149–60. https://doi.org/10.1007/978-3-642-21102-7.

R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2023.

Renfrew J, Sanderson H.Herbs and vegetables. In: Ghillean P, Mark N, editors. 2012.

RStudio Team. RStudio: Integrated Development for R. PBC, Boston, MA: RStudio; 2020.

Sastry E. Taramira (*Eruca sativa*) and its improvement–a review. Agric Rev. 2003;24(4):235–49.

Sihag RC. Pollination ecology of rocket (*Eruca vesicaria* (L.) Cav. ssp. *sativa* (Mill.) Thell) in the semi-arid environments of Northwest India: native bees are the major pollinators. Ecologies. 2023;4(3):580–94. https://doi.org/10.3390/ecologies4030038.

Subramanian B, Gao S, Lercher MJ, Hu S, Chen WH. Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. Nucleic Acids Res. 2019;47(W1):W270–5. https://doi.org/10.1093/nar/gkz357.

Sun W, Guan C, Meng Y, Liu Z, Zhang T, Li X, et al. Intergeneric crosses between *Eruca sativa* Mill. and *Brassica* species. Zuo Wu Xue Bao. 2005;1:36–42.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607–13. https://doi.org/10.1093/nar/gky1131.

Traboulay DM. Columbus and Las Casas: the conquest and Christianization of America, 1492–1566. Maryland: University Press of America; 1994.

Wang Y, Chu P, Yang Q, Chang S, Chen J, Hu M, et al. Complete mitochondrial genome of *Eruca sativa* Mill. (Garden rocket). PLoS One. 2014;9(8):e105748. https://doi.org/10.1371/journal.pone.0105748.

Warwick SI, Gugel RK, Gómez-Campo C, James T. Genetic variation in *Eruca vesicaria* (L.) Cav. Plant Genet Resour. 2007;5(3):142–53. https://doi.org/10.1017/S1479262107842675.

Xu G, Zhang X, Chen W, Zhang R, Li Z, Wen W, et al. Population genomics of *Zea* species identifies selection signatures during maize domestication and adaptation. BMC Plant Biol. 2022;22(1):72. https://doi.org/10.1186/s12870-022-03427-w.

Yaniv Z, editor. Tradition, uses and research on rocket in Israel. In: Rocket: a Mediterranean crop for the world. Report of a workshop. Legnaro (Padova); 1996.

Zafar-Pashanezhad M, Shahbazi E, Golkar P, Shiran B. Genetic variation of *Eruca sativa* L. genotypes revealed by agro-morphological traits and ISSR molecular markers. Ind Crops Prod. 2020;145: 111992. https://doi.org/10.1016/j.indcrop.2019.111992.

Publisher's Note

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