Analysis of >3400 worldwide eggplant accessions reveals two independent domestication events and multiple migration-diversification routes

Lorenzo Barchi[1,*] iD, Giuseppe Aprea[2], M. Timothy Rabanus-Wallace[3,4], Laura Toppino[5], David Alonso[6], Ezio Portis[1], Sergio Lanteri[1], Luciana Gaccione[1], Emmanuel Omondi[7], Maarten van Zonneveld[7], Roland Schafleitner[7], Paola Ferrante[2], Andreas Börner[3], Nils Stein[3,8] iD, Maria José Díez[6], Veronique Lefebvre[9] iD, Jérémy Salinier[9,10], Hatice Filiz Boyaci[11], Richard Finkers[12,13], Matthijs Brouwer[12], Arnaud G. Bovy[12] iD, Giuseppe Leonardo Rotino[5], Jaime Prohens[6] and Giovanni Giuliano[2,*] iD

[1]DISAFA − Plant Genetics, University of Turin, Grugliasco, Torino 10095, Italy,

[2]ENEA, Casaccia Res Ctr, Via Anguillarese 301, Rome 00123, Italy,

[3]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, Seeland, OT Gatersleben 06466, Germany,

[4]Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, Australia,

[5]CREA, Research Centre for Genomics and Bioinformatics, Via Paullese 28, Montanaso Lombardo LO 26836, Italy,

[6]Universitat Politècnica de València, Camino de Vera 14, Valencia 46022, Spain,

[7]The World Vegetable Centre, Tainan City, Taiwan,

[8]Department of Crop Sciences, Center for Integrated Breeding Research (CiBreed), Georg-August-University, Von Siebold Str. 8, Göttingen 37075, Germany,

[9]INRAE, GAFL, Montfavet F-84140, France,

[10]CIRAD La Réunion et Mayotte, UMR PVBMT Saint-Pierre, La Réunion, France,

[11]Department of Horticulture, Faculty of Agriculture, University of Recep Tayyip Erdogan, Rize, Turkey,

[12]Wageningen University & Research WUR, Wageningen, The Netherlands, and

[13]GenNovation B.V., Wageningen, The Netherlands

SUMMARY

Eggplant (*Solanum melongena*) is an important Solanaceous crop, widely cultivated and consumed in Asia, the Mediterranean basin, and Southeast Europe. Its domestication centers and migration and diversification routes are still a matter of debate. We report the largest georeferenced and genotyped collection to this date for eggplant and its wild relatives, consisting of 3499 accessions from seven worldwide genebanks, originating from 105 countries in five continents. The combination of genotypic and passport data points to the existence of at least two main centers of domestication, in Southeast Asia and the Indian subcontinent, with limited genetic exchange between them. The wild and weedy eggplant ancestor *S. insanum* shows admixture with domesticated *S. melongena*, similar to what was described for other fruit-bearing Solanaceous crops such as tomato and pepper and their wild ancestors. After domestication, migration and admixture of eggplant populations from different regions have been less conspicuous with respect to tomato and pepper, thus better preserving 'local' phenotypic characteristics. The data allowed the identification of misclassified and putatively duplicated accessions, facilitating genebank management. All the genetic, phenotypic, and passport data have been deposited in the Open Access G2P-SOL database, and constitute an invaluable resource for understanding the domestication, migration and diversification of this cosmopolitan vegetable.

Keywords: *Solanum melongena*, Single Primer Enrichment Technology, domestication, single nucleotide polymorphism, passport data.

INTRODUCTION

Solanum melongena L., known as eggplant, brinjal, or aubergine, is the third most cultivated Solanaceous crop species after potato (*S. tuberosum* L.) and tomato (*S. lycopersicum* L.), with a worldwide production exceeding 58.6 M tons in 2021 (FAO). Closely related species include the wild ancestor *S. insanum* L., and other species such as *S. incanum* L., *S. lichtensteinii* Willd., and *S. linnaeanum* Hepper & P.-M.L. Jaeger (Acquadro et al., 2017; Knapp et al., 2013; Page et al., 2019; Vorontsova et al., 2013). Different hypotheses on the origin of domesticated eggplant have been formulated: demographic studies suggest that, in contrast to tomato and pepper, eggplant did not undergo strong genetic bottlenecks during domestication (Arnoux et al., 2021). Most studies agree on the 'out of Africa' hypothesis, with some of them suggesting that the African species *S. incanum* spread to Asia, where it was domesticated through the intermediate species, *S. insanum*. Historical evidence points to early cultivation of eggplant in India and Southern China. Some studies identified a single domestication center in Southeast Asia and hypothesized that the Indian *S. insanum* accessions were, in fact, *S. melongena* that reverted to feral, weedy forms (Page et al., 2019; Weese & Bohs, 2010), while others have identified at least two domestication centers, in Southeast Asia and India (Meyer et al., 2012). A comparison of the medicinal uses of eggplant in several regions of Asia supports the hypothesis of multiple domestication centers (Meyer et al., 2014). Linguistic and historical evidence points to both India and China as ancient cultivation hotspots: the names aubergine (English–French), berenjena (Spanish), and al-bāḏinjān (Arabic) derive from the Sanscrit vātiṃgaṇa, pointing an Indian early cultivation; and the earliest written record of eggplant cultivation dates back to 59 B.C. in China (Daunay & Janick, 2007; Wang et al., 2008). *S. melongena* seeds have been found in mineralized plant remains from an excavation of an Abbasid Jerusalem bazaar, dating back to the 8th-9th century A.D. (Fuks et al., 2020).

Additional domesticated species are the scarlet eggplant (*S. aethiopicum* L. gr. *gilo*) and the gboma eggplant (*S. macrocarpon* L.), mainly cultivated in the African continent, whose respective wild ancestors are *S. anguivi* Lam. and *S. dasyphyllum* Schumach. & Thonn. A more distant group comprises the American eggplant relatives (Syfert et al., 2016; Vorontsova et al., 2013), among which *S. torvum* Sw. and *S. sisymbriifolium* Lam. are potentially exploitable for eggplant breeding, due to their resistance to several important eggplant diseases (Daunay & Hazra, 2012) (Figure 1a).

About 7.4 million accessions of crops and wild relatives are currently maintained in over 1700 genebanks worldwide (Harrison, 2017). Genebanks represent valuable phenotypic data providers, since a number of highly heritable traits, routinely scored during each multiplication cycle, have a potential use for genome-wide association studies (Milner et al., 2019). However, sharing germplasm among genebanks frequently results in corruption or loss of the associated metadata, leading to inaccurate classification and/or unintentional duplicates (Milner et al., 2019). Large-scale genotyping represents a valuable tool for identifying putatively duplicated samples with distinct information and correcting taxonomic misassignments (Langridge & Waugh, 2019; Milner et al., 2019; Tripodi et al., 2021). Among the technologies used for genotyping, Single Primer Enrichment Technology (SPET) (Barchi et al., 2019; Herrero et al., 2020; Scaglione et al., 2019; Villanueva et al., 2021) represents a customizable, cost-efficient solution, but
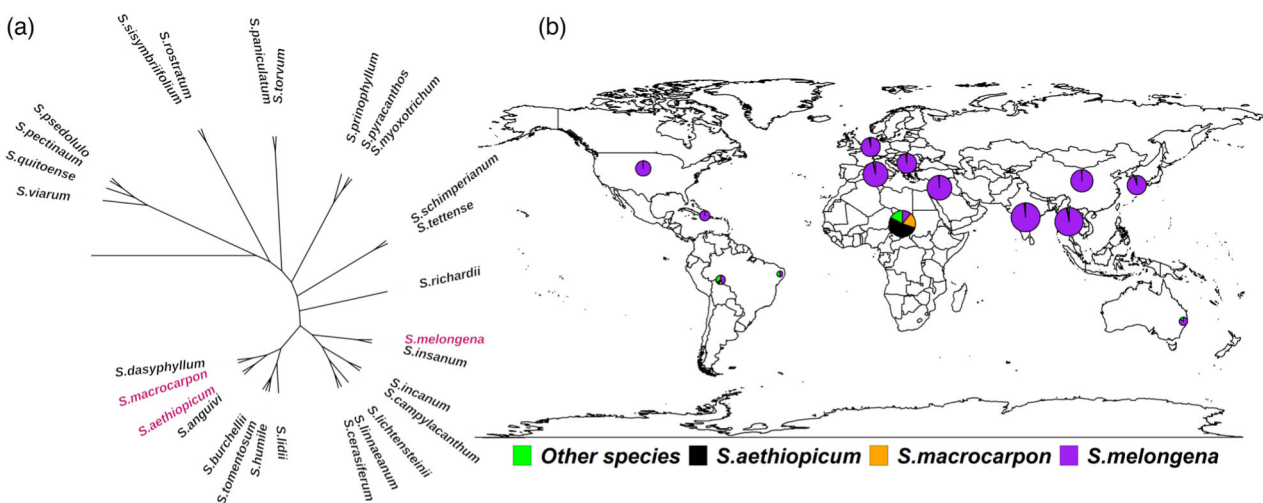
Figure 1. Species relationships and geographical origin of the different accessions included in the study.
(a) Cladogram of the *Solanum* taxa used in this work based on Särkinen et al. (2013) and Vorontosova et al. (2013); domesticated species are in magenta.
(b) Geographical origin of the accessions. The area of each pie chart is proportional to the number of accessions from a specific geographical area, while the color refers to the species.

requires *a priori* genomic or transcriptomic information and identification of single nucleotide polymorphisms (SNPs) for probe design. High-quality eggplant genome sequences have recently become available (Barchi et al., 2021; Barchi et al., 2019; Li et al., 2021; Wei et al., 2020), allowing the development of a panel of 5K SPET probes for eggplant (Barchi et al., 2019).

Here, we report the constitution and SPET-based genotyping of a worldwide collection of 3499 accessions of domesticated eggplants and 25 wild relatives, conserved in seven international genebanks and research institutions. Based on the passport, geographical origin, and genomic data, we evaluated the population structure of the species and identified mislabeled accessions and putative duplicates within and among *ex-situ* collections, greatly facilitating the genebank management workflows. Furthermore, these analyses allowed a reexamination of the possible domestication centers and diffusion routes of *S. melongena*, as well as the characteristics selected by different local user groups.

RESULTS AND DISCUSSION

A worldwide collection of *S. melongena* and its relatives

A total of 3499 accessions of eggplant and its relatives (Figure 1a) held by seven international genebanks were identified, and their passport data were collected from genebank archives and manually curated for homogeneity (see 'Experimental procedures' section). Based on this information, the accessions originated from 105 countries across five continents. The largest part of the collection analyzed included accessions from the Indian subcontinent (24.8%), Southeast Asia (21.5%), and Africa (14.0%) (Figure 1b; Table S1). A total of 3335 accessions belonged to domesticated species (*S. melongena*, 2896 accessions; *S. aethiopicum*, 313; and *S. macrocarpon*, 123 accessions) of which 3 were *S. melongena* hybrids with other species, while the remaining 164 belonged to 25 wild species (Table S2).

SPET genotyping of the collection

The accessions were subjected to SPET genotyping using a previously developed 5K probes panel (Barchi et al., 2019). About 500K trimmed, quality filtered reads per accession (average coverage 85.3×) were produced and aligned to the reference '67/3' eggplant genome v4.1 (Barchi et al., 2021) with an average mapping rate of 95% (Table S1). Three thousand four hundred twelve accessions, plus 31'67/3' controls showing a coverage of 10× or higher were used for the subsequent analyses.

After applying quality filters, a total of 119 695 polymorphic sites were identified among the 3412 accessions, of which 4306 were the target SNPs used for construction of the initial panel, while the remaining 115 389 were non-target SNPs found in flanking regions (Barchi et al., 2019;

Herrero et al., 2020; Mangino et al., 2022; Villanueva et al., 2021). The percentage of missing data in *S. melongena* was extremely low (0.03%) and increased with phylogenetic distance from *S. melongena*, from 0.10% in the direct progenitor of brinjal eggplant, *S. insanum*, to >15% in *S. torvum* and *S. sisymbriifolium*, which are native to the New World, up to 49% in *S. prinophyllum* Dunal, from Australia (Acquadro et al., 2017; Vorontsova et al., 2013) (Tables S1 and S3). The cultivated African species (*S. aethiopicum* and *S. macrocarpon*) showed missing data values of 0.25% and 0.79%, respectively, confirming the applicability of the 5K panel for the genotyping of most species in the eggplant genepool (Figure S1a; Table S3).

The average heterozygosity estimated using total SNPs (Figure S1b; Table S3) was very low for *S. melongena* (0.30%), its direct progenitor *S. insanum* (0.13%), and in the sister species *S. incanum* (0.26%), as well as for the two other cultivated species, that is, *S. aethiopicum* (0.66%) and *S. macrocarpon* (0.41%), indicating essentially autogamous reproduction. The higher average level of heterozygosity detected in the other species is attributable to a higher rate of allogamy (Daunay et al., 2001).

Identification of putative duplicates and taxonomic misassignments for genebank management

Sharing of germplasm among genebanks and manual record-keeping used in the past increased the chances of losing or misidentifying the metadata associated with each accession, thus resulting in inaccurate classification and/or unintentional duplication of accessions. To identify taxonomically misclassified accessions and potential duplicates, we estimated Identity-By-State (IBS) proportions of all their pairwise combinations, based on the genotype matrix of all the 119 695 polymorphic sites. The 31 replicate samples of *S. melongena* cv. '67/3' (Figure S2a) represented the cut-off value of dissimilarity (<0.0004) for identifying putative replicates (Figure S2b).

A total of 591 accessions (excluding the controls) were classified as potentially duplicated according to these criteria; they clustered in 212 groups, each including from two to 13 duplicates (Figure 2a). The highest number of duplications was detected for *S. melongena* (425 accessions), followed by *S. aethiopicum* (105 accessions) and *S. macrocarpon* (43 accessions). A variable number of intra-genebank putative duplicates (autoduplicates) were found in different genebanks, ranging from 0.0% at BATEM to ~23.6% at UPV (Table 1). Putative duplicates between different genebanks (alloduplicates) were also found: for instance, WorldVeg and INRAE shared 59 alloduplicates, while INRAE shared with CGN and UPV 36 and 30 alloduplicates, respectively (Table 1). In total, ~83% of accessions were unique according to the above criteria.

Most of the time, putative duplicates display very similar phenotypes (Figure 2b, b1, b3) but sometimes they
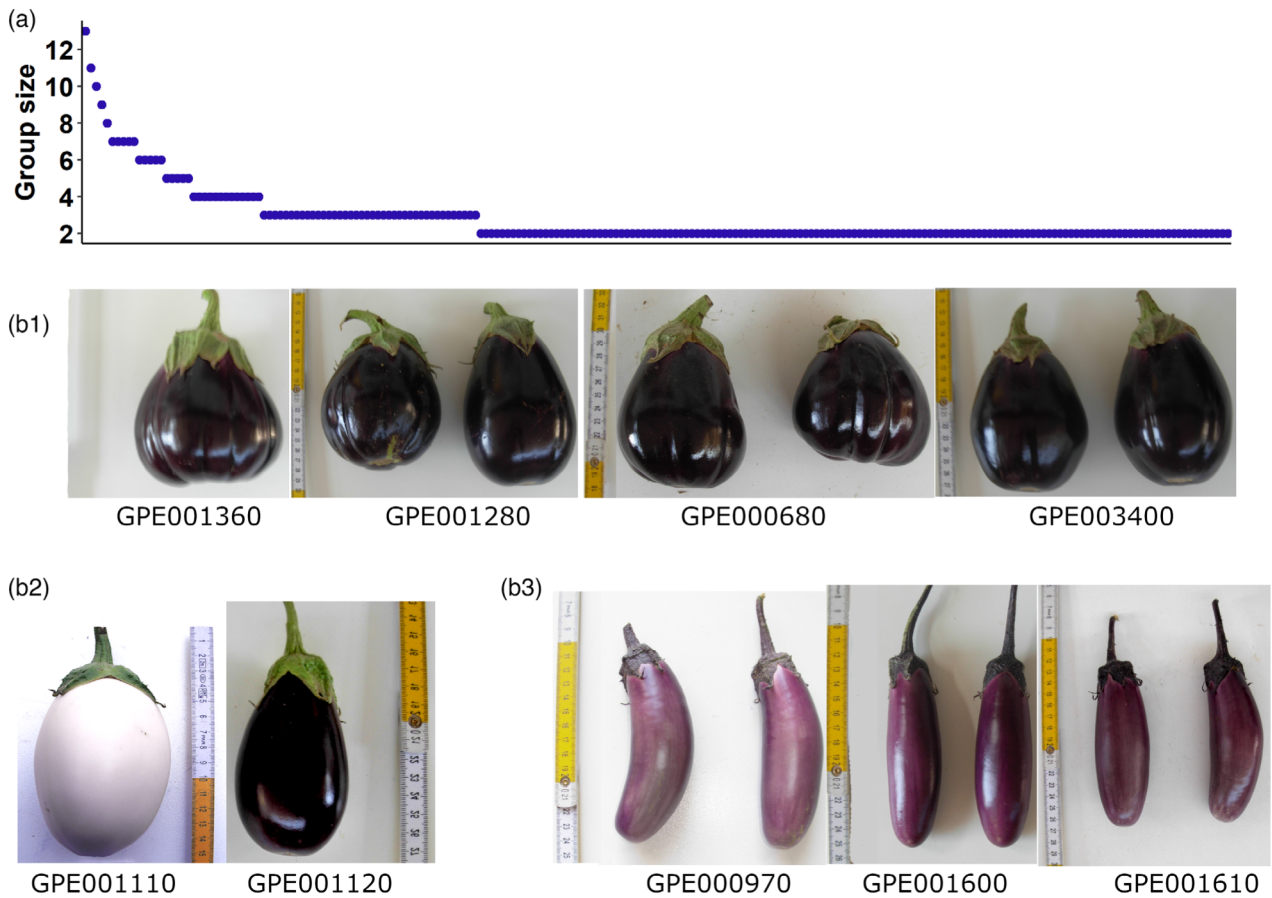
Figure 2. Putatively duplicated groups in genebank accessions.
(a) Number of accessions identified within each group of duplicates (reported on the *x* axis).
(b) Representative phenotypes of putatively duplicated accessions: (b1, b3) Same fruit typology and color; (b2) Similar fruit typology, but different color.

may also be characterized by different key agronomic or morphological traits (Figure 2b,b2). Since the SPET panel used was designed to study 5K random different genomic regions, single gene mutations resulting in large phenotypic differences may not be inspected by the panel. This observation highlights the central role of phenotypic observations by genebank curators in complementing the genotyping results; however, the genotyping approach provides a robust method and, in the case of unreliable or missing passport data, the only method for identifying potential duplicates.

To correct species assignments, we applied a 'nearest neighbor' analysis: groups of accessions with a dissimilarity of 0.0004 were inspected, and accessions were called as potentially misclassified if their species match rate with the 10 nearest neighbors was less than 35%. Overall, 88 accessions were identified as putatively mislabeled by this method (Tables S2 and S4). For 31 of them, it was possible to perform the correct assignment to a different species after the inspection and manual curation by genebank curators of the genebanks' passport and phenotypic data.

Study of intra- and interspecific phylogenetic relationships using different SNP datasets

The phylogenetic relationships among the 3412 eggplant accessions were explored using both the whole and target SNP datasets.

The Maximum Likelihood (ML) tree based on the whole SNP panel was in good agreement with previous reports (Figure S3a) (Acquadro et al., 2017; Barchi et al., 2019; Gramazio et al., 2017). Two main branches were identified, of which one included the species *S. prinophyllum* and a representative of *S. aethiopicum*. In the second branch, two clusters are recognizable: (i) accessions from the New World species *S. sisymbriifolium* and *S. torvum* as well as those belonging to the 'Lasiocarpa' clade; (ii) species native to the Old World. The latter cluster includes four sub-clusters: (i) representatives of the Madagascar species *S. myoxotrichum* Baker, *S. pyracanthos* Lam. and of the east African *S. tettense* Klotzsch and *S. schimperianum* Hochst. ex A. Rich.; (ii) two species of the 'Anguivi' grade (*S. macrocarpon* and its wild

Table 1 Percentage (%) of putatively duplicated accessions within (autoduplicates) and between (alloduplicates) genebanks

Autoduplicates	Count	Percentage
BATEM	0	0
CGN	48	11.2
CREA	49	21.03
INRAE	57	8.53
IPK	9	9.09
UPV	45	23.56
WorldVeg	160	9.08

Alloduplicates	Count	Percentage genebank 1	Percentage genebank 2
CGN-CREA	17	3.98	7.30
CGN-INRAE	36	8.43	5.39
CGN-IPK	18	4.22	12.12
CGN-UPV	12	2.81	4.19
CGN-WorldVeg	34	7.96	2.10
CREA-INRAE	19	8.15	2.10
CREA-IPK	7	3.00	5.05
CREA-UPV	11	4.72	7.33
CREA-WorldVeg	27	11.59	1.14
INRAE-WorldVeg	59	8.83	4.65
IPK-INRAE	8	8.08	0.60
IPK-WorldVeg	8	8.08	0.23
UPV-INRAE	30	15.71	7.34
UPV-IPK	4	2.09	2.02
UPV-WorldVeg	13	6.81	0.68

species belonging to the 'Eggplant' clade (*S. melongena*, *S. insanum*, *S. incanum*, *S. lichtensteinii*, *S. campylacanthum* Hochst. ex A. Rich., *S. cerasiferum* Dunal and *S. linnaeanum*).

Few accessions (11 out of 3412) did not fall within the expected clusters; since the passport data confirmed their original species assignment and the whole SNP dataset used provides enough resolution to discriminate between different species, a likely explanation of these discrepancies might be a mix-up of samples during DNA preparation and/or their management.

In the principal component analysis (PCA) (Figure 3a), the first and second components accounted for 3.85 and 2.99% of the whole genetic variation, respectively. The two American species *S. sisymbriifolium* and *S. torvum*, as well as those belonging to the 'Lasiocarpa' clade, were separated from the other accessions by the first component, while those of the 'Eggplant' clade and 'Anguivi' grade by the second one. With the exclusion of *S. campylacanthum*, the accessions closest to *S. melongena* belong to the 'Eggplant' clade (*S. insanum*, *S. incanum*, *S. lichtensteinii*, *S. cerasiferum* and *S. linnaeanum*), followed by the 'Anguivi' grade species (*S. aethiopicum* and *S. macrocarpon*).

A second ML tree and PCA analysis, based on target SNPs, maximized separation within *S. melongena* (Figure S3b; Figure 3b). The ML tree confirmed a clear separation among species, and two main branches were identified: (i) *S. melongena* accessions together with some *S. insanum* representatives; (ii) all the other species. The first and second components of the PCA accounted for 14.6 and 7.3% of the whole genetic variation, respectively, highlighting considerable
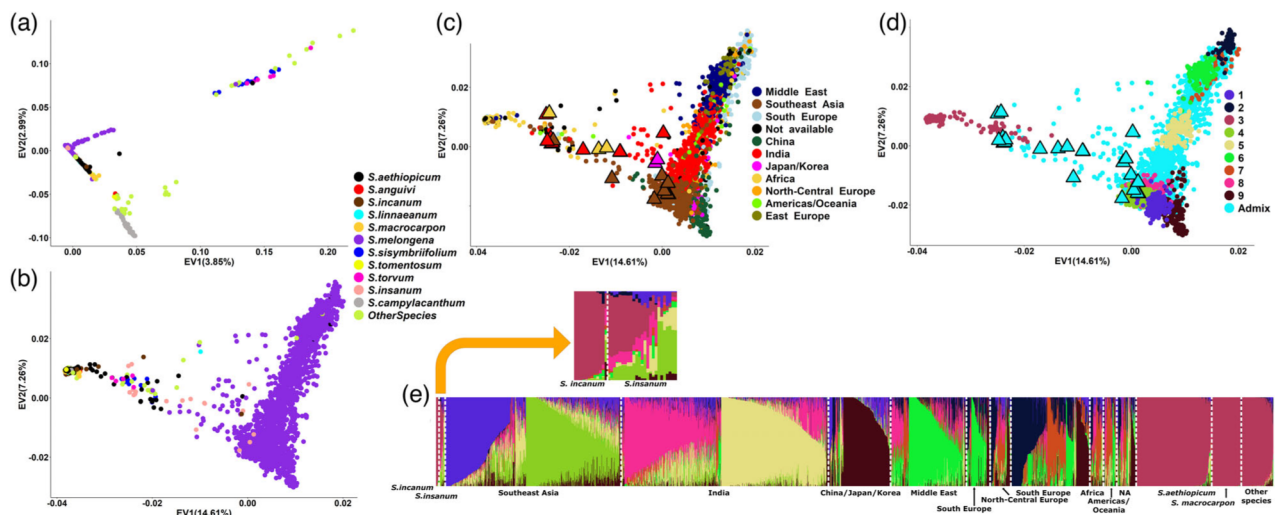
progenitor *S. dasyphyllum*); (iii) other representatives of the 'Anguivi' grade (*S. aethiopicum* and its wild progenitor *S. anguivi*), *S. humile* Lam. and *S. tomentosum* L.; (iv)

Figure 3. Principal component analysis (PCA) of eggplant genetic diversity in worldwide genebank holdings.
(a) PCA colored by species, whole single nucleotide polymorphisms (SNPs).
(b) PCA colored by species, target SNPs.
(c) PCA colored according to geographic origin, target SNPs.
(d) PCA colored according to Admixture *K* clusters, target SNPs (*K* = 9).
(e) Admixture clustering analysis, *K* = 9, target SNPs, using the same color scheme of (d). *Solanum insanum* and *S. incanum* admixture assignment are expanded in the panel indicated by the arrow. The full Admixture analyses, with *K* = 2–15, are shown in Figure S4.

differences with respect to the total SNP dataset. The first component separated *S. melongena* from the other species (Figure 3b) and spread over a large area. The rest of the 'Eggplant' clade and the 'Anguivi' grade species clustered with just minor overlapping. In agreement with previous findings on a smaller dataset (Barchi et al., 2019), when using the target SNPs, the American species *S. sisymbriifolium* and *S. torvum* plotted in an intermediate area between *S. melongena* and the other Old-World species.

Importantly, target SNPs allowed a clear separation of *S. melongena* accessions with respect to their geographical origin: eggplant accessions originating from Southeast Asia and the Indian subcontinent formed two main clusters in both the ML tree (Figure S3b) and PCA analysis (Figure 3c), each of which included several sub-groups. Accessions from China, Japan and Korea were well-separated from the others in both PCA and ML tree, as were European and Middle East accessions, in which some accessions of possible Chinese and Japanese origin were found, presumably as a result of historical import of materials.

These results confirmed our previous observation that the use of the total SNP dataset is more appropriate for broad phylogenetic studies (Barchi et al., 2019) and that the three domesticated species (*S. melongena*, *S. aethiopicum* and *S. macrocarpon*) belong to clearly separate groups that probably diverged at similar times, although contrasting results have been reported on their relationships (Acquadro et al., 2017; Gramazio et al., 2017; Isshiki et al., 2008; Meyer et al., 2012; Sakata et al., 1991; Sakata & Lester, 1997; Vorontsova et al., 2013). Furthermore, *S. aethiopicum* and *S. macrocarpon* together with their wild ancestors showed some level of admixture, probably as a result of genetic exchange between the domesticated and their wild ancestor (Lester & Hasan, 1991; Lester & Niakan, 1986; Plazas et al., 2014; Syfert et al., 2016). Conversely, target SNPs boosted intra-specific discrimination within *S. melongena*, and correlated strongly with geographical origin.

Eggplant domestication and gene flow

Different approaches were used to infer the history, population structure and domestication of the eggplant gene-pool. First, we measured the nucleotide diversity (π) and the fixation index F_{ST}, which gives a measure of the differentiation between two populations, between *S. melongena* groups of different geographical origin, its progenitor *S. insanum* and sister species *S. incanum*, using the target SNP dataset (Table 2). As expected, the highest F_{ST} values among these three species were observed between *S. melongena* and *S. incanum*, with the latter species exhibiting an unusually low π; this could be due either to a narrow genetic basis of this species, or to a sampling bias of the *S. incanum* accessions used in this study. In contrast, *S. insanum* exhibited the highest π among all the groups/species in the study, suggesting a broad genetic basis, in

agreement with its derivation from a variety of different geographical areas (Table S1). The *S. melongena* groups exhibited variable π and F_{ST} values, as a result of the different degrees of admixture due to migration and interbreeding (Arnoux et al., 2021; Barchi et al., 2021). The two geographical groups showing the lowest F_{ST} values with *S. insanum* (0.23 and 0.26, respectively) were those from India and Southeast Asia, indicating a close genetic relationship with the progenitor species, as well as with the sister species *S. incanum*, typical of groups directly derived from domestication events (Cao et al., 2022). Both groups exhibited relatively high π values, in agreement with the hypothesis that eggplant domestication occurred in the absence of strong genetic bottlenecks (Arnoux et al., 2021). Furthermore, in the PCA analyses we observed an East to West geographical gradient: the African and Indian *S. insanum* accessions were well-separated from their *S. melongena* counterparts, while those from Southeast Asia were much less separated (Figure 4a; Figure S5), suggesting that reversion of *S. melongena* to weedy *S. insanum* is more likely to have occurred in Southeast Asia.

To infer the genetic ancestry of *S. melongena* groups of different geographical origins, we used ADMIXTURE (Alexander et al., 2009) to carry out a model-based clustering approach, using >0.70 membership proportion and target SNPs (Figure S4). The ΔCV curve showed several inflection points, at $K = 3$, 4, 5, 7, and 9 (Figure S4a). At $K = 5$ and more evidently at $K = 7$, the Southeast Asia accessions were clearly separated from Indian ones, while at $K = 9$ both groups separated into two similarly sized subgroups. We chose $K = 9$ as the highest K showing an inflection point in ΔCV curve as well as a clear differentiation between the Mid-Eastern and South European accessions (Figure 3e). This analysis confirmed that the Indian and Southeast Asia accessions contained strong genetic components derived from *S. insanum*, the red and ochre components being predominant in India, and the blue and green ones in Southeast Asia. This clear separation of the genetic ancestry of accessions from the two regions further indicates the existence of at least two, and possibly up to four, domestication centers, localized in India and Southeast Asia. In *S. melongena*, a slight separation of Indian from Chinese and Southeast Asian accessions was also described by Page et al. (2019) following STRUCTURE analysis. However, presumably due to the limited number of accessions in their study, they reported an admixture among these two groups with just a transition gradient between *S. insanum* and *S. melongena*.

The domestication and migration history were also explored using TreeMix (Pickrell & Pritchard, 2012), which investigates population splits and admixtures without *a priori* hypotheses on the presence or absence of gene flow. Initially, 1–10 migrations were tested, using *S. incanum* as an outgroup and grouping accessions according to their

Table 2 Genetic diversity (π) and F_{ST} values *Solanum melongena* from different geographical areas, its progenitor *S. insanum* and sister species *S. incanum*. For π, values are ranging from lowest (dark blue) to highest (dark red). For Fst, the orange intensity increases with the increase of the value.

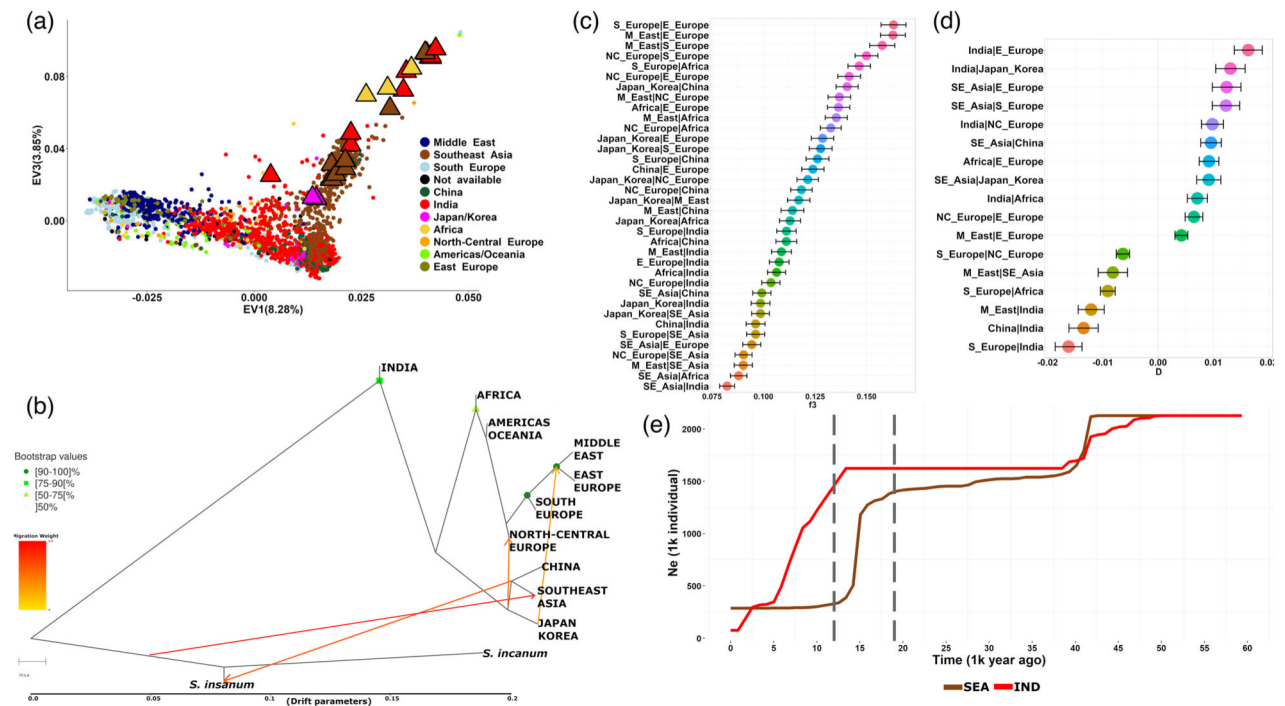| Origin	No. of accessions	Genetic diversity (π)	F_{ST}										
			Americas Oceania	East Europe	North Central Europe	South Europe	Middle East	Japan Korea	China	Southeast Asia	India	S. *insanum*	S. *incanum*
Africa	55	0.33	0.02	0.13	0.02	0.06	0.12	0.21	0.22	0.17	0.10	0.29	0.53
Americas and Oceania	55	0.32		0.14	0.04	0.07	0.10	0.24	0.26	0.20	0.11	0.31	0.54
East Europe	97	0.26			0.09	0.08	0.05	0.25	0.27	0.24	0.20	0.41	0.64
North and Central Europe	84	0.33				0.04	0.10	0.16	0.18	0.16	0.12	0.29	0.53
South Europe	322	0.28					0.08	0.24	0.25	0.24	0.18	0.38	0.59
Middle East	309	0.26						0.28	0.30	0.25	0.18	0.40	0.61
Japan and Korea	82	0.24							0.08	0.15	0.19	0.37	0.63
China	171	0.25								0.14	0.20	0.36	0.61
Southeast Asia	717	0.30									0.13	0.23	0.50
India	845	0.32										0.26	0.50
S. *insanum*	25	0.36											0.30
S. *incanum*	12	0.20											

Figure 4. Eggplant domestication from *Solanum insanum*.

(a) Plot of the first and third components from principal component analysis of *S. melongena* (dots) and *S. insanum* (triangles) using target single nucleotide polymorphisms according to their origin.

(b) Maximum likelihood (ML) tree with bootstrap support values with migration events from TreeMix, indicated by colored arrows. The color scale shows the migration weight. The scale bar shows 10 times the average standard error of the estimated entries in the sample covariance matrix.

(c) Outgroup f3-statistic for all possible admixture populations using *S. insanum* as outgroup. The higher the value, the more recently the two populations diverged.

(d) *D*-statistic to detect admixture according to W,X;Y,Z model. We tested admixture from *S. insanum* (Y) to either W (first population) or X (second admixture population), using *S. incanum* (Z) as outgroup. A negative *D*-statistic indicates that gene flow has occurred from *S. insanum* to X (second population in the names on the *y*-axis), and a positive *D* statistic indicates that gene flow has occurred from *S. insanum* to W (first population in the names on the *y*-axis). World regions of origin are coded as: ME: Middle East, SEA: Southeast Asia, SE: South Europe, CH: China, IND: India, AOC: Americas and Oceania, JK: Japan and Korea, AF: Africa, NCE: North and Central Europe, EE: East Europe.

(e) Stairway plot showing changes in effective population size (N_e) of *S. melongena* accessions grouped according to their origin (SEA: Southeast Asia, IND: India) through relative time. Gray dashed lines delimit the time of N_e decrease (12–19 kya).

geographical origin. To infer the best-supported migration model, we used the OptM package (Fitak, 2021), which indicated that four migrations were best supported (Figure S6a). A potential gene flow from *S. insanum* to Southeast Asian accessions was identified (Figure 4b), confirming the hypothesis of a primary domestication center in this region. Furthermore, an influx of alleles from Japanese and Korean accessions to the branching point between Middle East and East European accessions was identified, together with a gene transfer from 'Far East' accessions to North-Central European accessions.

Conversely, *S. insanum* seemed to harbor some genomic introgressions of brinjal eggplant from Southeast Asia and China, in line with previous findings (Page et al., 2019), a further evidence of the reversion of some Southeast Asian *S. melongena* to weedy *S. insanum*. Indeed, gene flow from domesticated species to their wild relatives can occur and be maintained for generations (Ellstrand et al., 2013). Page

et al. (2019) spotted pervasive gene flow responsible for admixture, suggesting that some wild accessions represented a hybrid swarm. Accordingly, the same phenomena were observed in the PCA (Figure 3d; Figure S5), with *S. insanum* accessions belonging to admixed group K = 9.

The matrix of residuals (Figure S6b) indicated how well the four model fits the data; positive residuals represent populations that are more closely related to each other than in the best-fit tree, thus candidates for admixture, while negative residuals indicate that a pair of populations is less closely related than resulting from the best-fit tree. Strongly positive residuals suggested candidate admixture events among South European, Japanese/ Korean and Chinese accessions.

We applied the outgroup f3-statistic – which measures the shared genetic drift between two populations relative to an outgroup, while F_{ST} measures lineage-specific genetic drift – to measure the shared drift between two

populations in the origin model, using *S. insanum* as an outgroup. As shown in Figure 4(c), accessions from Southeast Asia showed the lowest f3 value with respect to the other eggplant groups, suggesting that *S. melongena* was firstly domesticated in this region. The *D*-statistic based on *Z*-value and calculated as (pop1, pop2; *S. insanum*, and *S. incanum*) (Figure 4d), highlighted that gene flow was observed mainly from *S. insanum* to both Southeast Asian and Indian accessions, and to a lesser extent to Africa.

We also studied effective population size (N_e) dynamics, to derive insights into the impacts of past environmental factors and human domestication events. The N_e for Southeast Asia and India showed ancient decreases (40–45 kya) coinciding with the glacial maxima (Prell & Kutzbach, 1987) (Figure 4e). A second bottleneck was observed in both groups, starting earlier (around 16 kya) in Southeast Asia and later (around 13 kya) in India. The dating of these bottlenecks is compatible with the timing suggested for the domestication of eggplant (Arnoux et al., 2021). Interestingly, both the timing and the kinetics of the Southeast Asia and India bottlenecks are different, further reinforcing the independent domestication hypothesis.

The combination of the above analyses suggests that: (i) *S. insanum* is an intermediate species between *S. incanum* and *S. melongena*, showing strong admixture with the latter species, especially in Southeast Asia; (ii) eggplant domestication occurred in the absence of strong genetic bottlenecks; (iii) at least two domestication events can be identified: the first in Southeast Asia traced back to around 18 kya and the second, more recent (around 12 kya), in the Indian subcontinent, with limited genetic exchange between the two regions.

Post-domestication eggplant spread and diversification

The F_{ST} values (Table 2) suggested different spread routes from the initial domestication centers in India and Southeast Asia: the closest groups to the Indian accessions are those from Africa, the Americas and Oceania (presumably because they were introduced from different geographical origins by migrants), and North and Central Europe; those closest to the Southeast Asia accessions are China and Japan/Korea. The genetic diversity (π) values for the three former groups are higher than the latter two, suggesting a stronger genetic bottleneck in the 'out of Southeast Asia' migration than in the 'out of India' one. Accessions from Middle East, Southern and East Europe show a closer genetic relationship with Indian, than with Southeast Asian ones, and their π values are rather low.

The PCA analyses carried out using *S. melongena* and *S. insanum* confirmed the two hypothesized geographical spread routes ('out of Southeast Asia' and 'out of India') that eggplant faced during its post-domestication history (Figure S5a). Indeed, the genetic relationships between India and Africa/Europe on one side, and Southeast Asia and Japan/Korea on the other were confirmed by the first and second components of the PCA analyses (Figure S5a). The Chinese accessions were well separated from the Southeast Asian ones by the second component and showed a close relationship to Southeast Asia according to the first component (Figure S5a). The first component separation of the Chinese accessions corresponds to the predominance, in the Admixture analysis, of a genetic component in the Chinese accessions (marked in dark brown in Figure 3e) which is only present at very low frequencies in Southeast Asia and *S. insanum* and, to a lesser extent, in India.

The two main geographical spread routes were also suggested by the ML tree calculated using target SNPs on *S. melongena* and *S. insanum* accessions (Figure S3c). Indeed, the Southeast Asian accessions clustered in a main branch of the tree, adjacent to Chinese, Japan/Korea, and a group of Southern European accessions, while those from India and the majority of European accessions were in adjacent branches in a different tree position.

Interestingly, the Treemix analysis, as well as F_{ST} values and the PCAs (Figure S5a) suggest an additional and more nuanced 'out of Southeast Asia' eggplant spread route, which from Far East brought eggplant directly to Middle East and Europe. The f3-statistics (Figure 4b) also suggest that the spread of cultivated eggplant from Southeast Asia occurred in at least two separate events, compared to the 'out of India' route.

We also applied the updated version of the ReMIXTURE ('Regional Mixture') procedure recently applied to the study of pepper migrations by Tripodi et al. (2021). The data clearly confirm the observations reported above, that is, that Southeast Asian accessions contain considerable unique diversity, followed by Indian representatives (Figure 5a). They also show incomplete genetic overlap between the two regions, consistent with the existence of two independent centers of domestication. The distribution of IBS values reveals greater similarity within groups than between them as expected. In contrast with pepper, however, the difference is less pronounced (Tripodi et al., 2021), suggesting a lower degree of eggplant varietal exchange between regions (Table S5). This agrees with the comparatively greater eggplant F_{ST} values between regions (Table 2). These differences likely arise from the ease of transporting pepper in dry form and the cultural appeal of pungency which make peppers exceptionally amenable to long-distance trade (Tripodi et al., 2021).

Taken together, these analyses suggest that eggplant spread from Southeast Asia to Japan and China, as well as westward along the Silk Road into Western Asia, Europe, and Africa by Arab traders during the fourteenth century through the 'out of Southeast Asia' route, and then was introduced into the Americas soon after Europeans arrived there (Prohens et al., 2005) (Figure 5b). However, based on
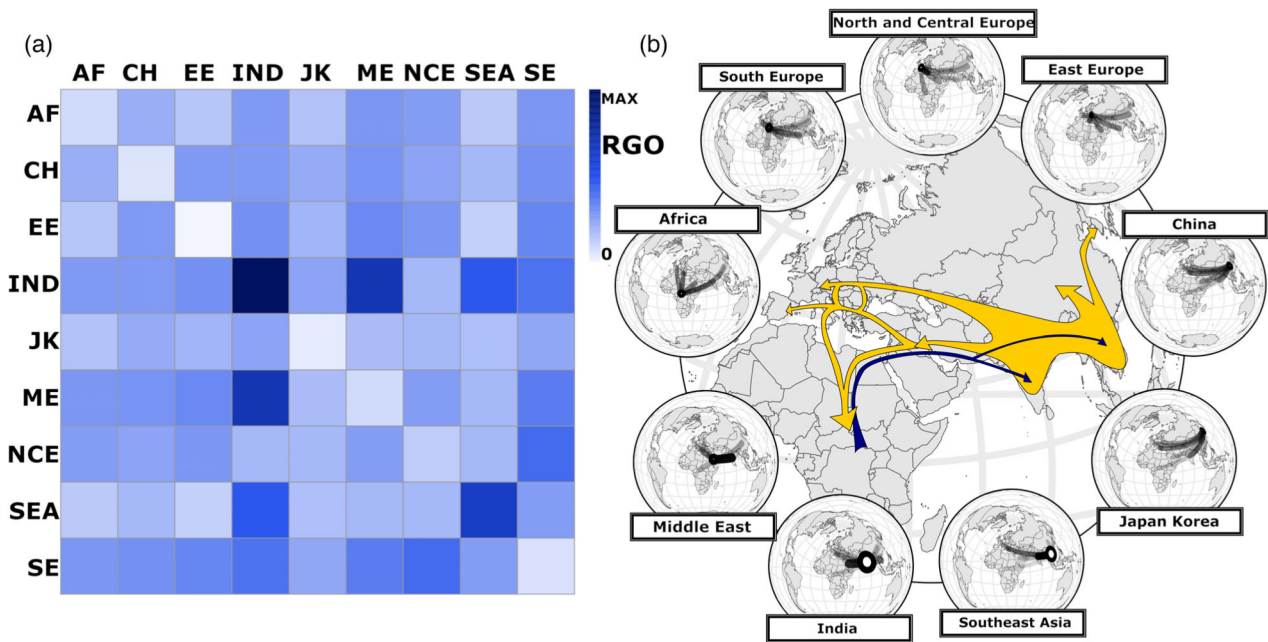
Figure 5. REMIXTURE analysis of eggplant dissemination from its domestication centers.
(a) Diversity matrix for 10 global regions. The relative amount of diversity that is uniquely possessed each region is shown on the diagonal, the amount of diversity overlapping the diversity of other regions off-diagonal. The highest diversity is found in the two proposed centers of domestication (Southeast Asia and the Indian Subcontinent), with modest overlap between the two.
(b) The major eggplant dissemination routes. In the regional maps in the periphery, the diameter of the black ring at the center of each region represents its total relative diversity, with the diameter of the white inner circle represents the diversity unique to that region. The thickness of lines joining regions indicates the total amount of overlapped diversity between regions, with the darkness of shading proportional to the amount relative to the total diversity of the focal region. The central cartoon summarizes the two domestication centers and the main exchange routes of *Solanum melongena*. Blue arrows indicate the 'out of Africa' hypothesis.

our findings, eggplant also dispersed from a second domestication center in the Indian subcontinent to Middle East, Africa, and Europe through the 'out of India' main spread route, in agreement with previous findings (Daunay, 2008; Prohens et al., 2005) as well as the historical evidences that Arab traders brought Indian eggplants into Europe, and Africa by the 14th century (Lewicki et al., 1974).

EXPERIMENTAL PROCEDURES

Germplasm collection

The accessions were provided by five international genebanks, including the World Vegetable Center (Tainan, Taiwan), the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK, Germany), the Universitat Politècnica de València Germplasm Bank (UPV-COMAV, Valencia, Spain), the Centre de Ressources Biologiques Légumes de l'Unité de Génétique et Amélioration des Fruits et Légumes (GAFL, INRAE, Montfavet, France), the Centre for Genetic Resources (CGN, Wageningen, The Netherlands), as well as by the genebanks of Universities and Research Institutions such as Batı Akdeniz Agricultural Research Institute (BATEM, Turkey) and Council for Agricultural Research and Economics (CREA, Italy) (Table S1).

DNA isolation and SPET library construction

DNA was extracted using the Qiagen plant mini-prep, the LGC Sbeadex kit, the SILEX protocol (Vilanova et al., 2020),

or a modified CTAB method, depending on the genebank. A total of 33 DNA samples of the reference *S. melongena* '67/3' line (nearly one per plate), obtained from a unique seed batch (Barchi et al., 2021; Barchi et al., 2019), were included as controls. The final set of 5082 (5K) probes previously identified was used and libraries were prepared as previously reported (Barchi et al., 2019) for genotyping the whole set of accessions at IGATech (Udine, Italy). Sequencing was performed on an Illumina NextSeq 500 platform (Illumina, Inc., San Diego, CA, USA), using 150SE chemistry. The raw sequencing data are available at NCBI SRA (BioProject ID PRJNA808188 and PRJNA542231). Accessions having an average read depth of <10 were discarded from the subsequent analyses (Table S1).

Read alignment and variant calling

Base calling and demultiplexing were carried out using the standard Illumina pipeline. The read quality check and adapter trimming were carried out using ERNE (Del Fabbro et al., 2013) and Cutadapt (Martin, 2011) software. After alignment to the reference eggplant genome (Barchi et al., 2021), using BWA-MEM (Li, 2013) with default parameters. SNP calling was obtained with GATK 4.1.9 (DePristo et al., 2011), following the software best practices in June 2021 for germline short variant discovery and as previously described in Barchi et al. (2019). To extract high-confidence SNPs, Vcftools (Danecek et al., 2011) was applied using the following parameters: min-meanDP 15 and no more than 5% of missing data.

Characteristics of the SPET panel used and use of different SNP datasets

The panel was designed using resequencing data from eight *S. melongena* and one *S. incanum* accessions (Gramazio et al., 2019), and thus the target SNPs were primarily intra-specific ones (Barchi et al., 2019). Given the small size of the group used for SNP discovery (the ascertainment group), the target SNPs are expected to be subject to ascertainment bias, for instance not containing rare SNPs that are monomorphic in the ascertainment group. Ascertainment bias is known to generate several artifacts with respect to non-biased methods such as whole genome sequencing (Malomane et al., 2018): an increase of expected heterozygosity, an underestimation of fixation index (F_{ST}) in populations with low (<0.15) F_{ST} values and overestimation of those with high (>0.15) F_{ST}. Various corrections can be applied to reduce the undesired effects of ascertainment bias, of which the most effective is linkage disequilibrium (LD-based SNP pruning; Malomane et al., 2018). In addition, since SPET is based on the sequencing of a 110-bp region surrounding the target SNP, it allows the discovery of several additional SNPs, not represented in the ascertainment panel, which are in LD with the target SNP. These 'non-target' SNPs increase in number proportionally to the genetic distance of the accessions genotyped. In fact, the heterozygosity of the *S. melongena* accessions was 0.30% when estimated using total SNPs, and 6.53% using target SNPs (Table S3), confirming the ascertainment bias of the target SNP panel used. As described previously (Barchi et al., 2019), using the total SNP dataset provided a good resolution between distantly related species, but very poor resolution of *S. melongena/S. insanum* accessions, which is instead maximized in the target SNP dataset (compare Figures 2 and 3 in this paper). Therefore, the choice of which SNPs dataset to use (total or target) was based on the purposes of the analysis (as described in the sections below). To minimize the effects of ascertainment bias for ML tree, PCA, Admixture, Treemix and f3-D statistics analyses, both datasets were pruned using plink2 (Chang et al., 2015) with the option '--indep-pairwise 50 10 0.2'.

Identification of sample duplication and misclassification

To estimate the degree of duplications and putative misclassifications, allele matching was calculated to provide an absolute percent IBS coefficient between all individuals. IBS was calculated using the snpgdsIBSNum function of SNPRelate (Zheng et al., 2012) using the full SNPs matrix. A distance-like matrix was then obtained by calculating the complement to 1 of the IBS matrix ($\Delta = 1 - IBS$) and all the nearest neighbor pairs were collected. The average distance between controls was used as a maximum threshold to define potential duplicates. For misclassification detection, for each accession, we determined the set of the 10 nearest neighbors and verified the species they belong to. A potential misclassification was called when the species match rate was less than 35%. The putative mislabeled accessions were eventually re-assigned to another species according to: (i) assigned to a group having at least six representatives of another species and (ii) after manual curation based on genebank passport data.

Genomic diversity and phylogeny

The analysis of genomic diversity and genetic relationships was inferred using several approaches. To gain a purely descriptive illustration of the genetic diversity in the samples, a PCA was performed with SNPrelate (Zheng et al., 2012) on both the total and target pruned SNPs datasets. Vcftools was used to calculate Weir and Cockerham's (1984) weighted F_{ST} and nucleotide diversity (π) using all target SNPs (Danecek et al., 2011) on *S. melongena*, *S. insanum* and *S. incanum* accessions.

We ran ADMIXTURE v.1.23 (Alexander et al., 2009) on both the total and target pruned SNPs datasets with the following parameters: number of subpopulations (K) ranging from K = 1 to 15, 20-fold cross-validation (CV). Each K was run with 20 replicates and the outputs were submitted to CLUMPAK (Kopelman et al., 2015). Individuals were tentatively assigned to one of the K populations if its membership coefficient (q_i) in that group was ≥0.70. We generated a dendrogram representation of the population's structure in the ML framework using IQ-TREE2 v2.1.3 (Minh et al., 2020) for both the total and target pruned SNPs datasets. Branch supports were obtained with the ultrafast bootstrap (Hoang et al., 2018) and the tree layout was generated using the online tool iTOL (http://itol.embl.de).

TreeMix 1.13 (Pickrell & Pritchard, 2012) was used on target SNPs to investigate population splits and admixtures without requiring prior hypotheses on the presence or absence of gene flow. We used OptM (Fitak, 2021) to estimate the optimal number of migration edges to add to the tree (from 1 to 10). BITE (Milanesi et al., 2017) was used to carry out 500 bootstrap replicates for the optimal migration events assumed and to visualize the consensus trees with bootstrap values and migration edges obtained by Phylip consense (Felsenstein, 2005).

Using *S. insanum* genotypes as an outgroup, we exploited outgroup f3-statistics (Patterson et al., 2012; Peter, 2016) calculated using ADMIXTOOL 2 (https://github.com/uqrmaie1/admixtools) on the target SNPs pruned (Maier et al., 2022) as a tool to measure shared drift between two populations according to geographical origin. To assess the direction of gene flow, we calculated D-statistics (Green et al., 2010) using ADMIXTOOL 2 on the target SNPs pruned. The D-statistics method considers the tree topology (((W, X), Y), Z) where Z represents the outgroup, Y the source of admixture, and W and X are the test populations. The D-statistics method counts the 'ABBA' sites, where W and Z share the outgroup allele (A), and X and Y share the derived allele (B), as well as the 'BABA' sites, where W and Y share the derived allele, and X and Z share the outgroup allele. Admixture between Y and either of the test populations creates a significant difference between the ABBA and BABA counts, with a z-score ≥3.0 (gene flow between W and Y) or ≤−3.0 (between X and Y).

The Python script easySFS was used to calculate the joint site frequency spectrum (SFS) for demographic analysis (https://github.com/isaacovercast/easySFS) using target sites in noncoding regions. Stairway Plot v.2 (Liu & Fu, 2020) was used to fit a multi-epoch demographic model to the data. A mutation rate of 2.35×10^{-8} per site per generation was used and a generation time was set to 1 year.

To elucidate the migration and sharing of eggplant varieties among different cultures, we used the updated version of the ReMIXTURE ('Regional Mixture') procedure (Tripodi et al., 2021) on the target SNPs. It begins by assigning a broad region of origin to each accession, and then clusters accessions using an UPGMA algorithm based on the IBS matrix. The number of clusters that feature a member of region A provide a proxy for the amount of genetic diversity a region A possesses. The proportion of these that also contain members of cluster B provides an analogous proxy for the proportion of A's diversity that is overlapped by B. The proportion of clusters comprising only individuals from A provides, similarly, a proxy for the proportion of A's diversity that is

unique. We refer to these proportions as self-overlap. Resampling is used to estimate standard errors around these values. The package's inbuilt parameter optimization tools (Figure S7) were used to select a resampling regime of 80% of accessions, resampled 5000 times, with a clustering cutoff of $h = 0.31$.

AUTHOR CONTRIBUTIONS

JP, SL, EP, NS, GLR, RS, LB and GG conceived the study. MTR-W performed the demography analysis. LT, DA, PF, AB, MJD, JS, VL, HFB, RF, MB, RF, AB, EO, RS and MvZ provided materials. LB, GA, LG produced and analyzed SPET data. LB and GG wrote the manuscript. All authors critically revised and approved the manuscript.

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the BioProject IDs: PRJNA542231 and PRJNA808188. The vcf containing the unfiltered SNPs is publicly available at https://doi.org/10.6084/m9.figshare.21716696. [Correction added on 14 October 2023, after first online publication: The link is updated in this version]. All the genetic, phenotypic and passport data are accessible at the G2P-SOL (http://www.g2p-sol.eu/G2P%2DSOL%2Dgateway.html).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Density plot of missing data (a) and heterozygosity (b) reported on the *x*-axis in the different species using all the SNPs dataset.

Figure S2. Controls used in the eggplant SPET genotyping. (a) Maximum likelihood tree region containing the 31 controls used. (b) Dissimilarity values of the control genotypes. Orange line highlights the cutoff value of 0.0004 for duplicate detection.

Figure S3. Maximum likelihood trees using the total SNP dataset (a), target SNPs (b), and target SNPs on *S. melongena* and *S. insanum* accessions (c). The colors in circle 1 correspond to species, in circle 2 to geographical origin (ME: Middle East, SEA: Southeast Asia, SE: South Europe, ND: Not Available, CH: China, IND: India, JK: Japan and Korea, AF: Africa, NCE: North and Central Europe, AOC: Americas and Oceania, EE: East Europe).

Figure S4. Admixture analyses using target (a) or total (b) SNPs datasets. Left: Delta of the cross-validation error estimate plot for the population structure. Right: model-based Admixture analysis with different numbers of ancestry kinship ($K = 2$–15). (A: *S. incanum*; B: *S. insanum*; C: *S. melongena* from Southeast Asia; D: *S. melongena* from India; E: *S. melongena* from China, Japan, and Korea; F: *S. melongena* from Middle East; G: *S. melongena* from East Europe; H: *S. melongena* from North central Europe; I: *S. melongena* from South Europe; L: *S. melongena* from North America; M: *S. melongena* Unknown; N: *S. melongena* from South American; O: *S. aethiopicum*; P: *S. macrocarpon*; Q: Other species).

Figure S5. PCAs of *S. melongena* (dots) and *S. insanum* (triangles) obtained using target SNPs. (a) Accessions are colored according to their origin. Left: first and second components, center: first and third components; right: second and third components. (b) Accessions are colored according to their admixture assignment. Left: first and second components, center: first and third components; right: second and third components.

Figure S6. Treemix analysis. (a) Optm output. (b) Heat map showing residuals. Zero is represented by white color. Residuals above zero represent pairs of populations that are more closely related to each other in the data than they appear in the best–fit tree and are, therefore, candidates for admixture events.

Figure S7. Optimization of the h-cutoff parameter for the ReMIXTURE procedure, showing the results of eight values of h run for 300 replicates each, each of which samples a random 80% of members of each region. Informative values for h produce enough clusters comprising members of each region to characterize the spread of values over replicate runs while balancing the relative amounts of mixed-region and single-region clusters. Values given on the y-axis are median cluster counts over the 300 replicates. Top: Optimal *h*-values tend to fall between the point of equal medians and the value that maximizes the number of multi-region clusters. Inspection of the heatmaps generated at each of these extremes shows no significant changes in the major trends noted. Bottom: The value of $h = 0.31$ chosen for h maximizes the multi-region cluster count while retaining enough clusters in each region.

Table S1. Characteristics, genebank propagation methods, and genotyping metrics of the 3499 accessions used in the study.

Table S2. Summary of the accessions in each species, before and after species correction.

Table S3. Missing data and heterozygosity calculated using total (top) and target SNPs (bottom).

Table S4. Mislabeled accessions identified.

Table S5. Top: eggplant mean inter-region IBS score. Bottom: pepper mean inter-region IBS score.

REFERENCES

Acquadro, A., Barchi, L., Gramazio, P., Portis, E., Vilanova, S., Comino, C. et al. (2017) Coding SNPs analysis highlights genetic relationships and evolution pattern in eggplant complexes. *PLoS One*, **12**, e0180774.

Alexander, D.H., Novembre, J. & Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

ADMIXTOOLS 2. Available at: https://github.com/uqrmaie1/admixtools. Accessed on 11th February 2023

Arnoux, S., Fraisse, C. & Sauvage, C. (2021) Genomic inference of complex domestication histories in three Solanaceae species. *Journal of Evolutionary Biology*, **34**, 270–283.

Barchi, L., Acquadro, A., Alonso, D., Aprea, G., Bassolino, L., Demurtas, O. et al. (2019) Single primer enrichment technology (SPET) for high-throughput genotyping in tomato and eggplant germplasm. *Frontiers in*

Plant Science, **10**, 1005. Available from: https://doi.org/10.3389/fpls.2019.01005/full

Barchi, L., Pietrella, M., Venturini, L., Minio, A., Toppino, L., Acquadro, A. *et al.* (2019) A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. *Scientific Reports*, **9**, 11769.

Barchi, L., Rabanus-Wallace, M.T., Prohens, J., Toppino, L., Padmarasu, S., Portis, E. *et al.* (2021) Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. *The Plant Journal*, **107**, 579–596.

Cao, Y., Zhang, K., Yu, H., Chen, S., Xu, D., Zhao, H. *et al.* (2022) Pepper variome reveals the history and key loci associated with fruit domestication and diversification. *Molecular Plant*, **15**, 1744–1758.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. & Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Daunay, M.-C. (2008) Eggplant. In: Prohens, J. & Nuez, F. (Eds.) *Vegetables II: Fabaceae, Liliaceae, Solanaceae, and Umbelliferae*. New York: Springer, pp. 163–220.

Daunay, M.-C. & Janick, J. (2007) History and iconography of eggplant. *Chronica Horticulturae*, **47**, 16–22.

Daunay, M.C., Lester, R.N., Gebhardt, C., Hennart, J.W. & Jahn, M. (2001) Genetic resources of eggplant (*Solanum melongena*) and allied species: a new challenge for molecular geneticists and eggplant breeders. In: van den Berg, R.G., Barendse, G.W.M., van der Weerden, G.M. & Mariani, C. (Eds.) *Solanaceae V: advances in taxonomy and utilization*. Nijmegen, The Netherlands: Nijmegen University Press, pp. 251–274.

Daunay, M.C.C. & Hazra, P. (2012) Eggplant. In: Hazra, P. & Peter, K.V. (Eds.) *Handbook of vegetables*. Houston: Springer, pp. 257–322.

Del Fabbro, C., Scalabrin, S., Morgante, M. & Giorgi, F.M. (2013) An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One*, **8**, e85024.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

easySFS. Available at: https://github.com/isaacovercast/easySFS. Accessed on 22 March 2023

Ellstrand, N.C., Meirmans, P., Rong, J., Bartsch, D., Ghosh, A., de Jong, T.J. *et al.* (2013) Introgression of crop alleles into wild or weedy populations. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 325–345.

FAO. Available at: http://faostat3.fao.org/home/E.org/. Accessed on 25 October 2022

Felsenstein, J. (2005) *PHYLIP (Phylogeny Inference Package)*. version 3.6.

Fitak, R.R. (2021) OptM: estimating the optimal number of migration edges on population trees using Treemix. *Biology Methods and Protocols*, **6**, bpab017.

Fuks, D., Amichay, O., & Weiss, E. (2020). Innovation or preservation? Abbasid aubergines, archaeobotany, and the Islamic Green Revolution. *Archaeological and Anthropological Sciences*, **12**, 50.

Gramazio, P., Prohens, J., Borràs, D., Plazas, M., Herraiz, F.J. & Vilanova, S. (2017) Comparison of transcriptome-derived simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers for genetic fingerprinting, diversity evaluation, and establishment of relationships in eggplants. *Euphytica*, **213**, 264.

Gramazio, P., Yan, H., Hasing, T., Vilanova, S., Prohens, J. & Bombarely, A. (2019) Whole-genome resequencing of seven eggplant (*Solanum melongena*) and one wild relative (*S. incanum*) accessions provides new insights and breeding tools for eggplant enhancement. *Frontiers in Plant Science*, **10**, 1220.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M. *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Harrison, R. (2017) Freezing seeds and making futures: endangerment, hope, security, and time in agrobiodiversity conservation practices. *Culture, Agriculture, Food and Environment*, **39**, 80–89.

Herrero, J., Santika, B., Herrán, A., Erika, P., Sarimana, U., Wendra, F. *et al.* (2020) Construction of a high density linkage map in oil palm using SPET markers. *Scientific Reports*, **10**, 9998.

Hoang, D.T., Chernomor, O., Haeseler, A.v., Minh, B.Q. & Vinh, L.S. (2018) UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, **35**, 518–522.

Isshiki, S., Iwata, N. & Khan, M.M.R. (2008) ISSR variations in eggplant (*Solanum melongena* L.) and related solanum species. *Scientia Horticulturae*, **117**, 186–190.

Knapp, S., Vorontsova, M.S. & Prohens, J. (2013) Wild relatives of the eggplant (*Solanum melongena* L.: Solanaceae): new understanding of species names in a complex group. *PLoS One*, **8**, e57039.

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A. & Mayrose, I. (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, **15**, 1179–1191.

Langridge, P. & Waugh, R. (2019) Harnessing the potential of germplasm collections. *Nature Genetics*, **51**, 200–201.

Lester, R. & Hasan, S. (1991) *Origin and domestication of the brinjal eggplant, Solanum melongena, from S. incanum, in Africa and Asia*. Kew, Richmond: The Royal Botanic Gardens.

Lester, R.N. & Niakan, L. (1986) Origin and domestication of the scarlet eggplant, *Solanum aethiopicum*, from *S. anguivi* in Africa. In: D'Arcy, W.G. (Ed.) *Solanaceae: Biology and systematics*. New York: Columbia University Press, pp. 433–456.

Lewicki, T., Johnson, M. & Abrahamowicz, M. (1974) *West African food in the Middle Ages: according to Arabic sources*. Cambridge: Cambridge University Press.

Li, D., Qian, J., Li, W., Yu, N., Gan, G., Jiang, Y. *et al.* (2021) A high-quality genome assembly of the eggplant provides insights into the molecular basis of disease resistance and chlorogenic acid synthesis. *Molecular Ecology Resources*, **21**, 1274–1286.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. Available from: http://arxiv.org/abs/1303.3997. Accessed on 13 June 2022

Liu, X. & Fu, Y.-X. (2020) Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biology*, **21**, 280.

Maier, R., Flegontov, P., Flegontova, O., Changmai, P. & Reich, D. (2022) On the limits of fitting complex models of population history to genetic data. *bioRxiv*. 2022.05.08.491072. Available from: https://doi.org/10.1101/2022.05.08.491072

Malomane, D.K., Reimer, C., Weigend, S., Weigend, A., Sharifi, A.R. & Simianer, H. (2018) Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC Genomics*, **19**, 22.

Mangino, G., Arrones, A., Plazas, M., Pook, T., Prohens, J., Gramazio, P. *et al.* (2022) Newly developed MAGIC population allows identification of strong associations and candidate genes for anthocyanin pigmentation in eggplant. *Frontiers in Plant Science*, **13**, 847789. Available from: https://doi.org/10.3389/fpls.2022.847789

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**(1), 200.

Meyer, R.S., Bamshad, M., Fuller, D.Q. & Litt, A. (2014) Comparing medicinal uses of eggplant and related Solanaceae in China, India, and The Philippines suggests the independent development of uses, cultural diffusion, and recent species substitutions. *Economic Botany*, **68**, 137–152.

Meyer, R.S., Karol, K.G., Little, D.P., Nee, M.H. & Litt, A. (2012) Phylogeographic relationships among Asian eggplants and new perspectives on eggplant domestication. *Molecular Phylogenetics and Evolution*, **63**, 685–701.

Milanesi, M., Capomaccio, S., Vajana, E., Bomba, L., Garcia, J.F., Ajmone-Marsan, P. *et al.* (2017) BITE: an R package for biodiversity analyses. *bioRxiv*. Available from: https://www.biorxiv.org/content/10.1101/181610v1 [Accessed 6th December 2021].

Milner, S.G., Jost, M., Taketa, S., Mazón, E.R., Himmelbach, A., Oppermann, M. *et al.* (2019) Genebank genomics highlights the diversity of a global barley collection. *Nature Genetics*, **51**, 319–326.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Haeseler, A.v. *et al.* (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, **37**, 1530–1534. Available from: https://doi.org/10.1093/molbev/msaa015

Page, A., Gibson, J., Meyer, R.S. & Chapman, M.A. (2019) Eggplant domestication: pervasive gene flow, feralization, and transcriptomic divergence. *Molecular Biology and Evolution*, **36**, 1359–1372.

Patterson, N., **Moorjani, P.**, **Luo, Y.**, **Mallick, S.**, **Rohland, N.**, **Zhan, Y.** *et al.* (2012) Ancient admixture in human history. *Genetics*, **192**, 1065–1093.

Peter, B.M. (2016) Admixture, population structure, and F-statistics. *Genetics*, **202**, 1485–1501.

Pickrell, J.K. & **Pritchard, J.K.** (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, **8**, e1002967.

Plazas, M., **AndÃojar, I.**, **Vilanova, S.**, **Gramazio, P.**, **Herraiz, F.J.** & **Prohens, J.** (2014) Conventional and phenomics characterization provides insight into the diversity and relationships of hypervariable scarlet (*Solanum aethiopicum* L.) and gboma (*S. macrocarpon* L.) eggplant complexes. *Frontiers in Plant Science*, **5**, 318.

Prell, W.L. & **Kutzbach, J.E.** (1987) Monsoon variability over the past 150,000 years. *Journal of Geophysical Research: Atmospheres*, **92**, 8411–8425.

Prohens, J., **Blanca, J.M.** & **Nuez, F.** (2005) Morphological and molecular variation in a collection of eggplants from a secondary center of diversity: implications for conservation and breeding. *Journal of the American Society for Horticultural Science*, **130**, 54–63.

Sakata, Y. & **Lester, R.N.** (1997) Chloroplast DNA diversity in brinjal eggplant (*Solanum melongena* L.) and related species. *Euphytica*, **97**, 295–301.

Sakata, Y., **Nishio, T.** & **Matthews, P.J.** (1991) Chloroplast DNA analysis of eggplant (*Solanum melongena*) and related species for their taxonomic affinity. *Euphytica*, **55**, 21–26.

Särkinen, T., **Bohs, L.**, **Olmstead, R.G.** & **Knapp, S.** (2013) A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. *BMC Evolutionary Biology*, **13**, 214.

Scaglione, D., **Pinosio, S.**, **Marroni, F.** *et al.* (2019) Single primer enrichment technology as a tool for massive genotyping: a benchmark on black poplar and maize. *Annals of Botany*, **124**(4), 543–552. Available from: https://doi.org/10.1093/aob/mcz054/5424191

Syfert, M.M., **Castañeda-Álvarez, N.P.**, **Khoury, C.K.**, **Särkinen, T.**, **Sosa, C.C.**, **Achicanoy, H.A.** *et al.* (2016) Crop wild relatives of the brinjal eggplant (*Solanum melongena*): poorly represented in genebanks and many species at risk of extinction. *American Journal of Botany*, **103**, 635–651.

Tripodi, P., **Rabanus-Wallace, M.T.**, **Barchi, L.** *et al.* (2021) Global range expansion history of pepper (*Capsicum* spp.) revealed by over 10,000 genebank accessions. *Proceedings of the National Academy of Sciences of the United States of America*, **118**, e2104315118.

Vilanova, S., **Alonso, D.**, **Gramazio, P.**, **Plazas, M.**, **García-Fortea, E.**, **Ferrante, P.** *et al.* (2020) SILEX: a fast and inexpensive high-quality DNA extraction method suitable for multiple sequencing platforms and recalcitrant plant species. *Plant Methods*, **16**, 110.

Villanueva, G., **Rosa-Martínez, E.**, **Şahin, A.**, **García-Fortea, E.**, **Plazas, M.**, **Prohens, J.** *et al.* (2021) Evaluation of advanced backcrosses of eggplant with *Solanum elaeagnifolium* introgressions under low N conditions. *Agronomy*, **11**, 1770.

Vorontsova, M.S., **Stern, S.**, **Bohs, L.** & **Knapp, S.** (2013) African spiny *Solanum* (subgenus *Leptostemonum*, Solanaceae): a thorny phylogenetic tangle. *Botanical Journal of the Linnean Society*, **173**, 176–193.

Wang, J.-X., **Gao, T.-G.** & **Knapp, S.** (2008) Ancient Chinese literature reveals pathways of eggplant domestication. *Annals of Botany*, **102**, 891–897.

Weese, T.L. & **Bohs, L.** (2010) Eggplant origins: out of Africa, into the orient. *Taxon*, **59**, 49–56.

Wei, Q., **Wang, J.**, **Wang, W.**, **Hu, T.**, **Hu, H.** & **Bao, C.** (2020) A high-quality chromosome-level genome assembly reveals genetics for important traits in eggplant. *Horticulture Research*, **7**, 1–15.

Weir, B.S. & **Cockerham, C.C.** (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

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