Classification: ARTICLE - Discoveries

**Eggplant domestication: pervasive gene flow, feralisation and transcriptomic divergence**

**Anna Page a, Jane Gibson a, Rachel S. Meyer b and Mark A Chapman a,\***

a Biological Sciences, University of Southampton, Southampton, SO17 1BJ, UK

b Department of Ecology and Evolutionary Biology, 610 Charles E. Young Drive South, Los Angeles, CA 90095, USA

\* Author for correspondence: Dr Mark A. Chapman; Phone: +44 (0)2380 594396; Email: [m.chapman@soton.ac.uk](mailto:m.chapman@soton.ac.uk); Orcid: 0000-0002-7151-723X

Keywords: Admixture; Aubergine; Breeding; Crop Evolution; Crop-wild gene flow; Domestication; Eggplant; Feral; Introgression

**ABSTRACT**

In the context of food security, examining the genomics of domestication will help identify genes underlying adaptive and economically important phenotypes, for example larger fruit, improved taste and loss of agronomically inferior phenotypes.

Examination of genome-scale single nucleotide polymorphisms demonstrates the relationships between wild ancestors of eggplant (*Solanum melongena* L.), confirming that *S. insanum* L. is the wild progenitor. This species is split roughly into an Eastern (Malaysian, Thai, and Vietnamese) and Western (Indian, Madagascan, and Sri Lankan) group, with domesticates derived from the former. Additional ‘wild’ accessions from India appear to be feral escapes, derived multiple times from domesticated varieties through admixture. Accessions with small egg-shaped fruit are generally found intermixed with East Asian *S. insanum* confirming they are primitive relative to the large-fruited domesticates.

Comparative transcriptomics was used to track the loci under selection. Sequence analysis revealed a genetic bottleneck reducing variation by almost 50% in the primitive accessions relative to the wild species and a further 10% in the landraces. We also show evidence for selection on genes with a role in response to wounding and apoptosis.

Genes showing a significant difference in expression between wild and primitive or between primitive and landrace genepools were mostly (> 75%) downregulated in the derived populations and enriched for gene ontologies related to defence, flowering, signalling, and response to biotic and abiotic stimuli.

This work reveals genomic changes involved in crop domestication and improvement, and the population genetics work explains why defining the eggplant domestication trajectory has been so challenging.

**INTRODUCTION**

The need to feed a growing population in the face of increased climate unpredictability makes it vital to identify the alleles, genes, cultivars and crops that confer yield, tolerance, food quality and safety traits ([Godfray, et al. 2010](#_ENREF_36); [Poppy, et al. 2014](#_ENREF_72)). The genetic changes underlying the traits that have been under human selection for millennia have great agronomical value. Their continued discovery requires true wild progenitors to be located and studied ([Burke, et al. 2007](#_ENREF_10); [Harlan 1992](#_ENREF_41); [Meyer and Purugganan 2013](#_ENREF_62)) because researchers and breeders often look to wild germplasm for novel alleles that can help crops grow on marginal land, provide pest resistance, phytonutrient qualities, or marketable traits ([McCouch, et al. 2013](#_ENREF_60); [Tanksley and McCouch 1997](#_ENREF_80)). These loci of interest are among the tens to thousands of loci under selection throughout domestication and crop evolution.

Comparisons between wild and domesticated germplasm assume the wild populations are ancestral, however for many crops and some domesticated animals, some of those perceived wild populations may be feral or admixed (Wang et al., 2017). These present complementary, but very different, opportunities to identify useful genomic loci: feral plants are natural quantitative trait experiments where, in just a few generations, wild adaptive phenotypes have re-emerged, potentially through introgression ([Ellstrand, et al. 2010](#_ENREF_29); [Gressel 2005](#_ENREF_40)). However, morphological diagnosis of truly wild from feral or admixed crop relatives is often nearly impossible, and has been extremely difficult with limited genetic information and limited sample sizes to appropriately compare populations. This is exacerbated by the observation that germplasm collections of crop wild relatives are typically severely underrepresented ([Hay and Probert 2013](#_ENREF_42)).

Further, actual domestication processes may have involved wild, feral, or hybrid plants at multiple points over time ([Allaby, et al. 2008](#_ENREF_2); [Arnold 2004](#_ENREF_5)). Similarly, domestication may involve semi-independent transitions from a wild to primitively domesticated species and further diversification and improvement into landraces and then cultivars. Exploiting loci that display signatures of selection within and between wild-primitive-domesticated (and domesticated-feral) populations requires a prerequisite: that a very complex puzzle of domestication and subsequent introgression is solved, which is especially daunting in older domesticates with broad geographic range. An excellent exemplar is seen in the old world nightshade crop, eggplant (*Solanum melongena* L.; Solanaceae).

This study was designed to use genome-scale data to resolve eggplant’s domestication history, to identify candidate targets of selection during domestication, and to characterize the role feral escapes have played in crop and crop wild relative diversity. Eggplant has been used as a food for over 2000 years in China ([Wang, et al. 2008](#_ENREF_84)) and for 4500 years in India, based on traces of *Solanum* in Harappan cooking vessels ([Kashyap and Weber 2013](#_ENREF_49)). It is widely grown throughout tropical and sub-tropical Asia, and owing to the abundance of land in Asia spanning along the same latitude, both domesticated and wild or weedy forms can have a broad distribution. It is also a crop where many theories of domestication have been proposed that can serve as testable frameworks for the genetic relationship between the different cultivated and naturally occurring eggplants in South, East, and SE Asia, yet studies so far have relied on only small numbers of DNA sequences or random markers ([Knapp, et al. 2013](#_ENREF_50); [Lester and Hasan 1991](#_ENREF_53); [Meyer, et al. 2012](#_ENREF_61); [Weese and Bohs 2010](#_ENREF_85)). Based on these studies the wild progenitor of eggplant is unanimously *S. insanum* L. a species which is sometimes split taxonomically into *S. insanum* F (true wild *S. insanum*) and *S. insanum* E (feral escapees; [Daunay, et al. 2001](#_ENREF_21)).

Eggplant is an old world crop, unlike its congeners tomato (*Solanum lycopersicum* L.) and potato (*S. tuberosum* L.), yet parallels exist, especially with tomato, in terms of the characteristics that were under selection. The genetic basis of domestication traits in eggplant might involve the same genes as in tomato based on comparative QTL (quantitative trait locus) analysis ([Doganlar, et al. 2002](#_ENREF_22)) and placement of eggplant orthologues of tomato domestication genes on a QTL map ([Portis, et al. 2014](#_ENREF_73)). Only a draft version of the ca. 1,127 Mb eggplant genome is currently available ([Hirakawa, et al. 2014](#_ENREF_43)).

We first carried out population genetic analysis using genome-wide single nucleotide polymorphisms (SNPs) to infer the evolutionary relationships of wild and domesticated eggplant. Included in our samples were a group of cultivated accessions known as *Solanum ovigerum*, characterised by small, usually white, egg-shaped fruits. *S. ovigerum* was postulated by Lester to be a primitive domesticate or landrace of *S. melongena* (Lester and Hasan, 1991), restricted geographically to the Centre of Origin (Vavilov 1922). We note the term ‘primitive’ has multiple definitions including ancestral, local, or having admixture from wild species (Zeven, 1988), and the definition for *S. ovigerum* remains unestablished, and that population genomic evidence may shed light on the genetic signature of a rare primitive crop. We then sequenced the transcriptomes of a geographically diverse panel and after assembling and annotating the transcriptome, carried out population genetic tests for selection as well as identified gene expression divergence between the wild and domesticated varieties.

**RESULTS**

We first confirmed that all species analysed were diploid by estimating genome size for a subset of accessions using flow cytometry (supplementary table S1, Supplementary Material online). All of the species analysed in this study were diploid (genome size 2.34 to 2.53 pg/2C) and a more distantly related species, *Solanum campylacanthum* Hochst., was determined to be tetraploid (4.72 to 4.85 pg/2C).

**A high quality eggplant SNP panel points to a single domestication and ongoing crop-wild gene flow**

We gathered genotyping-by-sequencing (GBS) data for 95 accessions split across four wild species and *Solanum melongena* (supplementary table S2, Supplementary Material online). Of these, 68 were alleged *S. melongena* landraces hailing from the primary geographies of cultivation in Asia where domestication or early cultivation had been postulated ([Knapp, et al. 2013](#_ENREF_50); [Lester and Hasan 1991](#_ENREF_53); [Meyer, et al. 2012](#_ENREF_61); [Weese and Bohs 2010](#_ENREF_85)). Included in this group were 13 eggplant cultivars falling under the classification of *S. ovigerum*. Landraces instead of commercial cultivars were selected to reduce the chance that breeding and wide hybridisation would confound our analysis of the crops. After quality trimming (see methods) 173.9M reads (1.83M ± 0.67M [SD] per individual) were retained. SNPs were identified across the panel, and those with no more than 20% missing data were retained, leaving 4880 SNPs.

Maximum likelihood (ML) phylogenetic analysis was performed using *Solanum lichtensteinii* Willd., *S. linnaeanum* Hepper & P.-M.L. Jaeger, and *S. incanum* L.as outgroup species (fig. 1*A*; a full version with individual accessions labelled is available as supplementary fig. S1, Supplementary Material online). *Solanum incanum* formed a monophyletic group sister to *S. insanum* and *S. melongena* confirming an African origin of the group and subsequent diversification in Asia ([Knapp, et al. 2013](#_ENREF_50); [Weese and Bohs 2010](#_ENREF_85)). *Solanum insanum* was split into two groups (one group contained two Indian accessions, the other, accessions from Madagascar to India to Thailand, Indonesia and Malaysia), which were sister with high bootstrap support to the remaining accessions. The remaining accessions then formed a gradation of wild (*S. insanum* from Malaysia and Thailand) and primitive (*S. ovigerum*) eggplants that were sister (with low bootstrap support) to the main *S. melongena* group. The gradation of wild and primitive accessions generally showed grouping by geography as opposed to taxonomy, but in some cases with low bootstrap support (fig. 1, inset). Analysis of the same dataset with SVDquartets (Chifman and Kubato 2014), which estimates the phylogenetic relationships under the coalescent, gave broadly similar results (supplementary fig. S1*B*, Supplementary Material online), but with some key differences. Firstly, four wild eggplants nested amongst the landraces in the ML tree are found at the base of the landrace clade in the SVDquartets analysis (see below), and secondly, some *S. melongena* accessions were found in different subclades of landraces in the two analyses.

Domesticated eggplant, *S. melongena*, is broadly split (with low bootstrap support) into a primarily Indian group (includes one sample from the Philippines and one Indian *S. ovigerum*) and an East Asian group (with the exception of containing one sample from India and one *S. ovigerum* of unknown provenance) in the ML analysis. This split is less clear in the SVDquartets analysis where there appear to be two groups of Indian accessions, and two of the East Asian accessions.

Analysis using STRUCTURE ([Falush, et al. 2003](#_ENREF_32)), which assigns proportional membership of each individual into each of *K* clusters (supplementary fig. S2, Supplementary Material online), supported two genetic clusters (fig. 1*B*; the outgroup taxa were excluded). Individuals present in the two *S. insanum* clades based on the ML phylogeny have highest membership to cluster 1 (>50% to 100%; red portion of bars in fig. 1*B*) with the *S. melongena* individuals corresponding to cluster 2 (blue portion of bars in fig. 1*B*). The STRUCTURE analysis for *K* = 5 showed some level of support (supplementary fig. S2, Supplementary Material online). Under this scenario *S. insanum* is split between two clusters (yellow and orange in fig. 1*C*) and the remaining three clusters roughly comprise (1) individuals in the gradation between *S. insanum* and *S. melongena*, (2) domesticated *S. melongena* from India, and (3) domesticated *S. melongena* from China and South-East Asia, although with clear admixture between the latter three groups (fig. 1*C*).

Using the result for *K* = 2, the individuals described above in the gradation between outgroup *S. insanum* and *S. melongena* show between 65 and 100% membership to cluster 2 (i.e. are mostly to all allied with the domesticated *S. melongena* samples). Using DIYABC (see methods) to test different hypotheses concerning the origin of the two largest and monophyletic groups of individuals in this gradation (green box, inset to fig. 1*A*), and their relationship to other taxa, suggests that these are the products of gene flow between wild (outgroup *S. insanum*) and domesticated eggplants (fig. 1*D*). Based on the proportional admixture estimated by DIYABC these groups show greatest contribution (91-93%) from the *S. melongena* group and a small proportion of ancestry (7-9%) from wild *S. insanum*.

The DIYABC analyses suggest point estimates for the origin of *S. melongena* (i.e. the split from *S. insanum*) 12,200 generations ago (IQR 7240-19,900; Fig 1*D*) and 8920 generations ago (IQR 8070-9520; Fig 1*E*). Although these numbers differ and the ranges are large (especially for the former), they overlap considerably. If we assume one generation per year, this suggests that cultivated eggplant originated ca. 10,000-9000 YBP.

**Feral eggplant arose through introgression between landrace and wild eggplant**

We found that four accessions named as *S. insanum*, and exhibiting wild-like traits,are nested within the domesticated eggplants with high bootstrap support in the ML analysis, but exhibit a position in the ‘gradation’ with SVDquartets analysis (supplementary fig. S1B, Supplementary Material online). These four show potential admixture (wild-domesticated) based on the STRUCTURE results (fig. 1*A*,*B*). This admixture is again backed up by DIYABC (using the ML topology to define the groups), but in this case the admixture estimated suggests a greater ancestry from wild than from the domesticated groups (60-69% wild: 31-40% domesticated; fig. 1*E*). These four samples were named *S. insanum* E according the seed source (M.-C. Brand-Daunay, pers. comm.), highlighting their weedy morphological characteristics.

Gene flow between domesticated and wild accessions was further supported by TreeMix (Pickrell and Pritchard 2012), which investigates historical admixture without the need to hypothesise on the presence or absence of gene flow. In our analysis (supplementary fig. S3, Supplementary Material online), in which we tested 1 to 10 migrations, the percentage of significant (p < 0.05) migrations was 100% for models of 1 to 3 migrations, but decreased at higher numbers of migrations. There was also a decrease in the rate at which the variance explained increased when the numbers of migrations modelled increased beyond 3, therefore the 3 migration model was identified as the most well supported. These three migrations resulted in the admixed genomes in the gradation of eggplants and in one of the pairs of feral accessions (supplementary fig. S3*C*, Supplementary Material online).

**Small fruit size in eggplant is multiply derived and may be ancestral or due to post-domestication introgression from the wild ancestor**

Taxonomists have named domesticated eggplants with small egg-shaped and -sized fruits and an often sprawling habit as a separate species, *S. ovigerum*. Our GBS SNP phylogenetic and population genetic evidence demonstrates that these characteristic eggplants are not monophyletic, with *S. ovigerum* accessions grouping with geographically proximal wild *S. insanum* or *S. melongena* accessions (fig. 1*A*, see above). Their phylogenetic position appears to indicate that these ‘primitive’ accessions are ancestral to the larger-fruited *S. melongena* but appear to also be admixed (see above), with a domesticated-like genome introgressed with a small proportion of *S. insanum* alleles.

Further, two accessions morphologically similar to *S. ovigerum* are closely related to, and nested within, the domesticated (and large-fruited) *S. melongena* group. This suggests that the small-fruited phenotype has originated multiple times from large-fruited varieties.

**Eggplant domestication was accompanied by a 45% reduction in genetic diversity and divergent selection at hundreds of loci**

RNAseq data were generated for 28 accessions, with an average of 18.34M ± 4.49M (SD) reads per individual after quality trimming (supplementary table S2, Supplementary Material online; raw data are available from the NCBI short read archive under BioProject ID PRJNA526115). Following normalisation and assembly using Trinity ([Grabherr, et al. 2011](#_ENREF_38)), lowly expressed transcripts were removed (see Methods), resulting in 172,464 transcripts and 104,402 components (loosely akin to genes) with an N50, mean and median contig length of 1918 bp, 1035 bp and 522 bp, respectively. 46.5% of transcripts had a significant (e-30) BLAST hit to tomato and 31.2% to *Arabidopsis.* The number of components is greater than the estimated number of genes in the eggplant genome (85,446; [Hirakawa, et al. 2014](#_ENREF_43)) indicating that the transcriptome is somewhat fragmented, a not uncommon finding in *de novo* RNAseq experiments ([Honaas, et al. 2016](#_ENREF_44)). After trimming to a single transcript per component (see methods) and identifying the polymorphic sites, 44,539 sites (from 11,159 loci) remained with a base call for all individuals.

Using the delineations from the GBS ML analysis (i.e. avoiding accessions where the taxonomic designation and the position in the ML phylogeny disagreed), our transcriptome sequencing covered *S. incanum* (n = 3), *S. insanum* (n = 4), *S. ovigerum* (n = 6) and *S. melongena* (n = 15) (arrows in fig. 1*A*). Genetic diversity and differentiation was examined for these four groups. The minimum number of individuals with an allele for the locus to be included was: *S. incanum* (all 3 individuals), *S. insanum* (n ≥ 3), *S. ovigerum* (n ≥ 3) and *S. melongena* (n ≥ 9).

Nucleotide diversity (π; [Nei and Li 1979](#_ENREF_68)) was significantly greater in the two wild species (*S. incanum* π = 0.000844 ± 0.000005 [SE]; *S. insanum* π = 0.000982 ± 0.000004) than in *S. ovigerum* (0.000528 ± 0.000004) and *S. melongena* (0.000479 ± 0.000004) based on paired t-tests for loci found in both populations being compared (all P < 0.001; fig. 2*A*). In addition nucleotide diversity was significantly lower in *S. melongena* than *S. ovigerum* (paired t-test, *t* = 29.55, P < 0.001). Based on these data, the transition from wild *S. insanum* to primitive eggplant (*S. ovigerum*) coincided with a loss of ca. 47.2% of nucleotide diversity, and the transition from primitive eggplant to domesticated landraces (*S. melongena*) coincided with a further loss of ca. 10.4% of nucleotide diversity. Roughly 45% of the loci were monomorphic in *S. ovigerum* (32,184/71,620), and 30% were monomorphic in *S. melongena* (21,384/72,319). Despite the smaller sample size (n = 4), a smaller proportion (22.1%; 15,177/68,840) were monomorphic in *S. insanum*.

Differentiation between *S. insanum* and *S. ovigerum* and between *S. insanum* and *S. melongena* was relatively high. Average Dxy was 0.0011 ± 0.0009 (SD) and average FST was 0.108 ± 0.165 between *S. insanum* and *S. ovigerum* and these values were 0.0011 ± 0.0009 and 0.122 ± 0.166 between *S. insanum* and *S. melongena* (fig. 2*B,C*). Differentiation between *S. ovigerum* and *S. melongena* was understandably lower (Dxy, 0.0007 ± 0.0008; FST, 0.066 ± 0.119; fig. 2*B,C*). We observed 1971 fixed sites in 1408 loci between *S. insanum* and *S. ovigerum* (out of 67,315 loci; i.e. 2.1% of loci had ≥ 1 fixed site). Between *S. ovigerum* and *S. melongena* there were only 29 fixed sites in 22 loci (out of 70,137 loci; i.e. 0.03% of loci had ≥ 1 fixed site).

**Analysis of gene expression evolution during domestication reveals consistent down-regulation and increased nucleotide divergence**

The analysis of expression divergence between leaves of the wild *S. insanum* and the landraces (i.e. loci showing potential expression divergence during eggplant domestication) revealed 2395 loci differentially expressed (DE) at FDR < 0.05. The majority (2082/2395; 86.9%) of the DE loci were down-regulated in the *S. melongena* landraceswith more than half of these loci exhibiting zero expression in the landraces (1155/2082; 55.5%). Included in the list of DE loci were several putative resistance proteins and genes potentially involved in plant architecture and development, including putative orthologues of cryptochrome 1a (involved in response to blue light) and a member of the TCP family of transcription factors with roles in leaf development in *Arabidopsis* ([Aguilar-Martinez and Sinha 2013](#_ENREF_1)).

The comparison between wild *S. insanum* and primitive/ancestral eggplant *S. ovigerum* revealed a smaller number of DE loci (937) and again the majority of loci (687/937; 73.3%) were again down-regulated in the derived population. The transition from *S. ovigerum* to landraces was associated with significant (FDR < 0.05) expression divergence of 722 transcripts. The majority (611/722; 84.6%) were down-regulated in the landraces, relative to *S. ovigerum*, with expression undetected in the landraces for 62.9% (380/611) of the down-regulated transcripts. Within the downregulated transcripts were several with putative functions related to resistance, for example Pto-like Serine/threonine kinases, CC-NBS-LRR resistance proteins and a putative member of the I2 family of resistance proteins in tomato responsible for resistance to *Fusarium oxysporum* ([Ori, et al. 1997](#_ENREF_71)).

For the non-monomorphic loci which had sufficient data (see above) we compared inter-population differentiation for the DE and non-DE loci. DE loci showed significantly (Mann-Whitney U-tests, all P < 0.0001) greater differentiation (both Dxy and Fst) than the remainder of the transcriptome (fig. 3). For example, median Dxy and Fst for the loci DE between *S. insanum* and the landraces was 0.00117 and 0.25397, respectively, whereas the value for the remainder of the transcriptome was 0.00081 and 0.04151, respectively.

The DE loci also exhibited a higher number of fixed SNPs between populations than the non-DE loci. For example, between *S. insanum* and the landraces there were 103 fixed sites from 2395 DE transcripts (0.043 per locus), whereas for the non-DE loci this value was (0.023 per locus). Similarly, between *S. insanum* and *S. ovigerum* there were 0.088 and 0.028 fixed SNPs per locus for the DE and non-DE loci, respectively. For the comparison between *S. ovigerum* and the landraces there were no fixed SNPs for the DE loci (however only 29 were found in total).

**Loci under selection point to flowering, hormone response and plant defence pathways as affected targets**

The analyses of population divergence were used to identify loci exhibiting increased differentiation during domestication and to identify Gene Ontology (GO; [Ashburner, et al. 2000](#_ENREF_6)) terms which were over-represented. We expect that the loci with the highest FST and Dxy (i.e. the top 5%) are enriched for loci exhibiting divergent selection. Population comparisons were made between *S. insanum* and *S. melongena*, between *S. insanum* and *S. ovigerum*, and between *S. ovigerum* and *S. melongena*.

We compared GO terms found in the top 5% tail of the distribution of Dxy with the GO terms present in the full transcriptome. After removing multiple hits to the same putative orthologue, *Arabidopsis* orthologues with GO terms ascribed were identified and GO analysis performed ([Du, et al. 2010](#_ENREF_24)). Loci showing the top 5% of Dxy between *S. insanum* and *S. melongena*, and/or between *S. insanum* and *S. ovigerum* (i.e. putatively involved in eggplant domestication), showed a significant (FDR < 0.05) over-representation of loci involved in the GO terms “nucleosome assembly”, “programmed cell death”, “oxidoreductase activity”, and “symplast”, along with associated terms (supplementary table S3*A,B*, Supplementary Material online). GO analysis of those loci in the top 5% of the distribution of Dxy between *S. ovigerum* and *S. melongena* (i.e. putative ‘improvement/diversification’ genes; supplementary table S3*C*, Supplementary Material online) revealed a significant over-representation of a range of GO processes, including “response to wounding”, and “transferase activity”.

The same analysis was applied to the loci identified as DE between populations. Of the 2395 and 937 loci DE between *S. insanum* and *S. melongena* and/or between *S. insanum* and *S. ovigerum*, 659 and 458 unique *Arabidopsis* hits were identified. GO analysis of these sets of loci revealed significant over-representation of GO terms related “pollen-pistil interaction”, “defense response”, “jasmonic acid biosynthetic process”, and “photoperiodism, flowering”, along with related processes and functions (supplementary table S3*D,E*, Supplementary Material online). For example, our evidence of gene expression divergence during domestication of putative orthologues of the genes *PR4* (*PATHOGENESIS-RELATED 4*), a gene encoding a potentially antifungal protein, and *DMR6* (*DOWNY MILDEW RESISTANT 6*), may be due to different selection pressures related to the biotic environment ([Mishina and Zeier 2007](#_ENREF_64); [Zeilmaker, et al. 2015](#_ENREF_89)). Similarly, we found evidence for expression divergence of orthologues of *GI* (*GIGANTEA*), and *ELF3* (*EARLY FLOWERING 3*), the protein products of which directly interact and are involved in photoperiodism and flowering ([Yu, et al. 2008](#_ENREF_88)).

Amongst the DE loci we also identified a putative orthologue of tomato *SELF PRUNING 6A*. Members of the *SELF PRUNING* (*SP*) gene family are responsible for aspects of plant architecture and the transition to flowering in tomato and pepper ([Elitzur, et al. 2009](#_ENREF_27)). Interestingly the *SP6A* gene in tomato contains a premature STOP codon in tomato and is not expressed, whereas the gene in the wild relative *S. pennellii* does not contain this premature STOP and is expressed ([Carmel-Goren, et al. 2003](#_ENREF_12)); this is the same pattern of expression we identified, with significantly greater expression in the wild eggplant (about 67-fold).

Finally, we identified putative orthologues of violaxanthin de-epoxidase, dihydroflavonol 4-reductase and anthocyanidin synthase, which were significantly down-regulated in domesticated eggplant relative to *S. insanum*. These genes are involved in pigment accumulation ([Goldsbrough, et al. 1994](#_ENREF_37); [Niyogi, et al. 1998](#_ENREF_69); [Saito, et al. 1999](#_ENREF_78)) and could be involved in the differences in leaf pigmentation between wild and domesticated eggplant leaves (*S. insanum* often has a purple midrib and purple spines; supplementary fig. S4, Supplementary Material online).

For the comparison of *S. melongena* and *S. ovigerum* only 240 unique *Arabidopsis* hits were identified and no significant GO terms were identified. Removing the FDR correction and setting the level of significance at *P* < 0.005 identified the GO terms “response to other organism”, and “response to salt stress” as over-represented (supplementary table S3*F*, Supplementary Material online). Our evidence of expression divergence of orthologues of *CERK1* (*CHITIN ELICITOR RECEPTOR KINASE 1*) and *ATHCHIB* (*ARABIDOPSIS THALIANA BASIC CHITINASE*), both of which are involved in resistance and immunity ([Miya, et al. 2007](#_ENREF_65); [Verburg and Huynh 1991](#_ENREF_81)), highlights that selection for alteration of these pathways could have taken place.

Although only leaf transcriptomes were analysed, hence expression of fruit development genes might not be expected, we also interrogated the transcriptome for orthologues of known tomato fruit development genes ([Chakrabarti, et al. 2013](#_ENREF_14); [Frary, et al. 2000](#_ENREF_35); [Rodriguez, et al. 2011](#_ENREF_76)). We identified putative orthologues of *FASCIATED*, *SUN* and *SlKLUH*, but not *OVATE* or *FW2.2*. None of these genes were in the top 5% for wild-landrace differentiation (*FASCIATED* was in the top 7%) or showed evidence for differential expression.

A comparison of categories (i.e. severity) of SNP mutations revealed little differences in the distribution of synonymous and non-synonymous SNPs (and the severity of the amino acid substitution) between either the highly differentiated and not highly differentiated SNPs (supplementary fig. S5*A*, Supplementary Material online) or between those segregating within the wild and landrace gene pools (supplementary table S5*B*, Supplementary Material online This could mean that a large proportion of the causative mutations are located outside the coding regions, for example in the *cis*-regulatory regions (e.g. promoters), which supports the observation of significant gene expression divergence.

**DISCUSSION**

Knowledge of the origin and diversification of crops is of interest from an evolutionary as well as an applied point of view ([Burke, et al. 2007](#_ENREF_10); [Purugganan and Fuller 2009](#_ENREF_74)). From an evolutionary standpoint we can learn about the strength and timing of selection, gene flow, genetic bottlenecks and the relative importance of standing variation or new mutations as well as that of selection on expression variation versus sequence variation. From an applied angle, we can identify progenitor species, novel alleles and understand the genetic basis of agronomic characters such as fruit size, shape and colour, seed shattering, and pest and pathogen resistance ([Meyer and Purugganan 2013](#_ENREF_62); [Ross-Ibarra, et al. 2007](#_ENREF_77)). In this investigation we analysed genome-wide SNP data to understand the demography of domestication and subsequent feralisation of eggplant.

In our investigation we find strong support for a single origin in South-East Asia of domesticated eggplant which was accompanied by a 47% loss of genetic diversity; a moderate domestication bottleneck compared to that experienced by rice (62%; [Caicedo, et al. 2007](#_ENREF_11)) and common bean (72%; [Bitocchi, et al. 2013](#_ENREF_8)), but stronger than that experienced by maize (33%; [Wright, et al. 2005](#_ENREF_87)) and sorghum (14%; [Casa, et al. 2005](#_ENREF_13)). Eggplants and their wild relatives are all strongly andromonoecious; therefore human selection is likely the major cause of the reduced variation we see. An origin early in the Holocene (ca. 10,000-9000 YBP) is inferred, although it is important to bear in mind that a number of evolutionary processes, including selection and admixture, can result in inaccurate estimators of divergence time.

Prior to the most recent taxonomic work ([Knapp, et al. 2013](#_ENREF_50)), the wild progenitor of eggplant, *S. insanum*, was often separated into a true wild eggplant and feral derivatives (e.g. [Daunay, et al. 2001](#_ENREF_21)) or even into two species (e.g. [Meyer, et al. 2012](#_ENREF_61)). Our work highlights how this uncertainty could have arisen: four accessions with wild-like traits (named *S. insanum*), are nested within the domesticated eggplants with admixed genomes. It appears that these wild-domesticated hybrid individuals have escaped cultivation through gene flow from the wild. A recent analysis of two major strains of weedy rice suggested de-domestication involved relatively few changes from the cultivated rice genome, with no obvious adaptive introgression from wild rice ([Li, et al. 2017](#_ENREF_56)). This suggests that there are many pathways to feralisation which may or may not involve introgression from the wild.

Further, the true wild progenitor appears to show some introgression from cultivated eggplant, especially in South-East Asia. Gene flow from crops to wild relatives is not uncommon and crop alleles may persists for many generations (reviewed in [Ellstrand, et al. 2013](#_ENREF_30)). Microsatellite studies have documented the presence of *S. melongena* alleles in Indian *S. insanum* ([Mutegi, et al. 2015](#_ENREF_66)) furthering the notion that crop-wild gene flow is ongoing. In our study, pervasive gene flow appears to have caused admixture, thus a subset of the wild accessions (inset fig. 1*A*) more precisely represent a hybrid swarm, as has been demonstrated recently in wild rice ([Wang, et al. 2017](#_ENREF_83)). Given this extensive hybridisation, it is not surprising that previous investigations using smaller numbers of genetic markers have come to contrasting conclusions surrounding the origin of domesticated eggplant ([Knapp, et al. 2013](#_ENREF_50); [Lester and Hasan 1991](#_ENREF_53); [Meyer, et al. 2012](#_ENREF_61); [Weese and Bohs 2010](#_ENREF_85)).

Increasing fruit size as well as increased diversity of fruit sizes are classic parts of the fruit crop domestication syndrome as pointed out by Darwin ([1868](#_ENREF_20)). We show that a specific group of cultivars with smaller egg-shaped fruits and a sprawling habit (*S*. *ovigerum*), popular in cuisines throughout the range of cultivation, with purported different culinary uses, and significantly different phytochemical properties from other Asian eggplants or *S. insanum* ([Meyer, et al. 2015](#_ENREF_63)), are likely representative of ‘primitive’ domesticates. These cultivars are both ancestral to *S. melongena* and carrying a small number of *S. insanum* alleles. The *S. ovigerum* accessions are primarily found in the gradation between *S. insanum* and domesticated eggplant and as such appear sister to the domesticates, however the non-monophyly of *S. ovigerum* suggests ongoing gene flow (from the wild into cultivated accessions). It is possible that egg-shaped and -sized fruits arose once and the trait was transferred through cross breeding, resembling the pattern in rice where the glutinous allele of *Waxy* arose in *japonica* accessions and has subsequently been introgressed into a subset of *indica* lines ([Olsen and Purugganan 2002](#_ENREF_70)). Alternatively *S. ovigerum,* and therefore theegg-shaped and -sized fruits, could be multiply-derived, potentially through gene flow from the wild into a cultivated background. A similar situation has arisen in maize where adaptive introgression from wild teosinte (the crop progenitor) has aided the expansion of the crop into the Mexican Highlands ([Hufford, et al. 2013](#_ENREF_45)). Adaptive introgression has been documented in several other crops ([Janzen, et al. 2019](#_ENREF_47)).

One goal of domestication research is to identify specific genes and alleles that show promise for future breeding and crop improvement. In our investigation we identified hundreds of candidate transcripts based on comparisons between wild and landrace populations, although we note that with the small sample size of true wild eggplants some follow-up is required. In addition, this will serve as an important SNP resource in the future.

The gene-by-gene sequence analysis (i.e. studying the transcriptome for signatures of selection) suffers from two potential shortfalls in the absence of a well-assembled genome. First, polymorphism was low in eggplant (45% and 30% of the loci were monomorphic in *S. ovigerum* and cultivated eggplant, respectively) and as such calculating the ratio of polymorphism in the wild vs. the domesticated populations was impossible for a large portion of the loci. Even if polymorphism in the cultivated population was changed to a very low but non-zero value, those loci with the ‘greatest reduction in diversity’ are simply those with the greatest polymorphism in the wild. Secondly, the physical linkage between loci on chromosomes would mean that any locus identified as having experienced a selective sweep, may well simply be in linkage disequilibrium (LD) with the actual target of selection. We therefore only identified putative targets of selection based on high differentiation between progenitor and derived populations, and not a drop in polymorphism; although we note that this still does not take into account the effects of LD. We also carried out the GO analysis on the high Dxy candidates and not the high Fst candidates because the latter statistic is artificially inflated when polymorphism is low in one or both populations under consideration ([Chapman, et al. 2016](#_ENREF_15); [Cruickshank and Hahn 2014](#_ENREF_19)), as was the case here.

Our analyses, based on both high Dxy and differential expression, identified thousands of loci with the potential signatures of divergent selection between wild-primitive and/or primitive-improved populations of eggplant. Although we expect some of these to be false positives, we found that a number of GO terms related to potential domestication-related phenotypes were over-represented. This highlights larger-scale selection patterns instead of identifying individual genes with potential evidence for diversifying selection. In particular it appears that genes involved in response to the biotic and abiotic environment, as well as hormone response and the regulation of flowering were under selection. Some overlap between our results and those from other crops is evident. For example, selection on stress response genes has been reported to have occurred during the domestication of common bean and sorghum ([Bellucci, et al. 2014](#_ENREF_7); [Mace, et al. 2013](#_ENREF_58)), selection on genes involved in response to auxin were identified in sorghum and maize ([Mace, et al. 2013](#_ENREF_58); [Wright, et al. 2005](#_ENREF_87)), and selection on genes involved in flowering time was reported as part of the domestication of common bean and sunflower ([Bellucci, et al. 2014](#_ENREF_7); [Chapman, et al. 2008](#_ENREF_17)).

The differential expression of a putative orthologue of tomato *SP6A* is particularly intriguing and may represent a case of parallel evolution. This gene is not expressed in cultivated tomato because of a premature stop codon. Comparative QTL analysis suggests some domestication traits are conferred by the same genomic regions in tomato, eggplant and pepper ([Doganlar, et al. 2002](#_ENREF_22); [Frary, et al. 2014](#_ENREF_34)), hence the same genes may be involved in parallel domestication, as has been shown for genes involved in shattering in cereal crops ([Lin, et al. 2012](#_ENREF_57)).

The increased Dxy and Fst in the DE loci, relative to the remainder of the transcriptome, suggests that diversifying selection on gene expression has been accompanied by selection at the sequence level too, potentially in the promoter region of the genes. Although significant portions of upstream regions may be missed from RNAseq investigations, the physical linkage to the coding region, would likely result in selection on the promoter region being evidenced as selection on the coding region. This increase in sequence differentiation has also been demonstrated in an investigation of DE between two wild plant species ([Chapman, et al. 2013](#_ENREF_16)). We also note that the molecular basis of expression divergence during domestication is, in the majority (86.9%) of cases, due to down-regulation or silencing of expression in the landraces, a pattern shared with common bean ([Bellucci, et al. 2014](#_ENREF_7)), but not with maize ([Swanson-Wagner, et al. 2012](#_ENREF_79)).

**CONCLUSIONS**

The longstanding sympatry of crops and their wild relatives, along with some level of reproductive compatibility, may mean that many crops exist as part of a wild-weedy-domesticated complex with occasional or even persistent gene flow (de Wet and Harlan, 1975; [Ellstrand 2003](#_ENREF_28); [Ellstrand, et al. 2010](#_ENREF_29); [Ellstrand, et al. 1999](#_ENREF_31); [Zizumbo-Villarreal, et al. 2005](#_ENREF_90)). Eggplant certainly seems to fit into this category with evidence for introgression and the transfer of adaptive alleles between wild and domesticated populations. This highlights the potential for crop-wild gene flow to be undetected without genome-wide analysis, and could potentially lead to the origin of feral weeds, if it hasn’t already.

Not only does this have implication for weed evolution (e.g. economic loss, [Ellstrand, et al. 2010](#_ENREF_29)) and for the potential release of transgenic crops, including eggplants ([Jayaraman 2010](#_ENREF_48)), this scenario also represents an exciting and tractable model to study adaptive evolution and the origin and transfer of adaptive traits in parallel. Once the eggplant genome becomes available, it will be possible to detect the size of introgressed regions, answering whether this is historic admixture (such as between Neanderthals and humans, [Vernot and Akey 2014](#_ENREF_82)) or recent and continual crop-wild gene flow (e.g. [Arias and Rieseberg 1994](#_ENREF_4); [Mandel, et al. 2016](#_ENREF_59)). This will also be fruitful in identifying candidate genes within introgressed regions that may correspond to adaptive phenotypes for a food secure future ([Godfray, et al. 2010](#_ENREF_36); [Poppy, et al. 2014](#_ENREF_72)).

**MATERIALS AND METHODS**

***Germplasm investigated***

The domestication of eggplant may have taken place in more than one geographic location; hence we sampled wild and landrace accessions from throughout the range as well as several outgroup species (supplementary table S2, Supplementary Material online). We broadly separated the domesticated accessions into three geographic groups following ([Meyer, et al. 2012](#_ENREF_61)), as follows: Indian accessions (including the Maldives), mainland Southeast Asian (comprising China, Taiwan, Thailand and Vietnam) and accession from the Malay Archipelago (Philippines, Malaysia and Indonesia). The subspecies of cultivated eggplant with characteristic egg-shaped fruit, *S. ovigerum*, was also sampled (supplementary table S2, Supplementary Material online). Although tissue for RNA and DNA was sampled from seedlings (see below) we grew the plants to maturity to ensure the accessions were morphologically typical of their designation, and in three cases we made changes to their grouping (supplementary table S2, Supplementary Material online).

Seed were mostly obtained from INRA (http://institut.inra.fr/en), with some from the AVRDC (http://avrdc.org/) and from personal collections (L. Bohs, University of Utah and R.S. Meyer). Seed were germinated on damp filter paper in a growth room with 16h daylength and temperature of 23°C. After germination, seedlings were transferred to pots of 3:1 Levington’s M2+S: vermiculite mixture in a glasshouse with 16h daylength supplemented by fluorescent bulbs.

For the population genomic analysis we samples 95 individuals (supplementary table S2, Supplementary Material online). For RNAseq we sequenced a smaller number of accessions, which were selected based on the results of the population genomic analyses.

***Genome size estimation***

Genome size was estimated for six of the sequenced accessions plus another four related accessions (in total: three each *S. campylacanthum* [not analysed in this study] and S. melongena, two *S. incanum* and one each *S. insanum* and *S. lichtensteinii* Willd.). Fresh leaves were harvested and sent overnight on wet ice to Plant Cytometry Services (http://www.plantcytometry.nl/). 2C values were calculated using flow cytometry based on comparison to the internal standard *Pachysandra terminalis* Siebold & Zucc. (3.5 pg/2C).

***DNA and RNA extraction and sequencing***

DNA was extracted using a modification of the ([Doyle and Doyle 1990](#_ENREF_23)) CTAB extraction protocol. Briefly, samples were homogenised and incubated for one hour at 60°C with 700ul of 2% CTAB buffer. Homogenate was cleaned twice using chloroform:Iso-amyl alcohol (24:1) and DNA precipitated using isopropanol, re-suspended in TE, and treated with RNase at 37°C for one hour. Samples were then precipitated with ethanol and re-suspended in water. Sample quantity was estimated using a Nanodrop1000 (NanoDrop Products, Wilmington, DE, USA) to ensure enough DNA was present for the sequencing. Samples were then sent to the Genomic Diversity Facility at Cornell University for GBS. Samples were digested using *Pst*I and sequenced using the standard Cornell GBS pipeline. After de-multiplexing, raw reads were analysed on the University of Southampton Iridis4 supercomputer.

Seedling tissue was used for RNA extraction because of the small amount of morphological differentiation between taxa at this young stage (compared to increased distinction as the plants mature). This should reduce the false identification of differential expression in loci which are influenced by varietal differences between the plants and was employed in an analysis of gene expression in tomato ([Koenig, et al. 2013](#_ENREF_51)).

After three weeks of growth in the glasshouse, one fully expanded leaf from each accession was removed, placed in a microfuge tube and immediately frozen in liquid nitrogen. From this, RNA was extracted using a Qiagen RNeasy Plant Mini kit (Qiagen, UK), utilising an on-column DNase step as per the manufacturer’s (Qiagen) recommendation. RNA quantity was approximated using a NanoDrop1000 (NanoDrop Products) to ensure high enough yields, whereupon the samples were sent to the Wellcome Trust Centre for Human Genetics (WTCHG, Oxford, UK) for accurate quantification and quality checking using a Tapestation (Agilent Technologies, Santa Clara, CA) and subsequent library preparation using Illumina’s Stranded Truseq kit (Illumina, UK). Libraries were sequenced across three lanes of Illumina Hiseq2000 for 101 cycles (paired-end), with up to 12 libraries (individually barcoded) combined per lane. Following the run, samples were de-multiplexed according to unique barcodes using the WTCHG/Illumina bioinformatic pipeline. The rest of the bioinformatic analyses took place as follows on the University of Southampton Iridis4 supercomputer.

***GBS analysis***

Raw reads from the 95 individuals were processed and assembled without a reference genome in ipyrad v.0.7.28 ([Eaton 2014](#_ENREF_26)). First, ipyrad de-multiplexes and filters the raw reads. Of the 192,843,367 raw reads, 18,906,611 total reads were removed based on missing RAD-tags (56,930), ambiguous barcodes (18,635,398), and using quality score filtering (214,283), retaining 173,936,756 reads.

Following filtering, ipyrad clusters reads based on similarity within samples, then estimates consensus sequences for these clusters. These consensus sequences are then clustered between samples to form loci, and filtered. The threshold for clustering at both stages is controlled by parameter 14 (p14), the percentage similarity at which two sequences are identified as homologous. After trying different settings, we chose p14 = 0.925 in an effort to prevent paralogous loci from merging incorrectly, as well as to ensure true alleles collapsed into a single locus. Parameter 24 (p24) defines the maximum number of shared polymorphic sites in a locus, to remove erroneously clustered paralogues. We used p24 = 0.25 which allowed heterozygous site to occur across a maximum of 25% of samples. Following this, loci were retained only if 80% of samples had data at that locus (parameter 21 = 76).

A maximum likelihood tree was constructed with RAxML, as part of the ipyrad analysis toolkit. Smart Model Selection (SMS) was used to determine the most appropriate evolutionary model. GTR+G+F - where GTR is the substitution rate matrix, and +G and +F are decorations - was selected using the Akaike information criterion. 500 bootstrap replicates were used to calculate branch support. SVDquartets (Chifman and Kubato 2014) within PAUP (<http://paup.phylosolutions.com/>) was used to estimate the phylogenetic relationships under the coalescent.

The same SNP dataset was used in a STRUCTURE analysis ([Falush, et al. 2003](#_ENREF_32)). The length of the burn-in period was 20,000, followed by 50,000 MCMC replicates. K values 1-8 were iterated over 10 times. Structure harvester ([Earl and von Holdt 2012](#_ENREF_25)) was used to find the most strongly supported K. The 10 replicates for each K value were aligned using CLUMPP ([Jakobsson and Rosenberg 2007](#_ENREF_46)).

Examination of alternate demographic scenarios was made for two specific subsets of the data; the first referring to the potential admixture of the ‘gradation’ of *S. insanum* and *S. ovigerum* accessions, and the second to the origin of feral eggplants (see Results). We employed DIYABC ([Cornuet, et al. 2014](#_ENREF_18)) for this analysis and tested multiple scenarios (fig. 1). The summary statistics used to derive the model were *proportion of zero values* and *mean of compete distribution* for genic diversities, Fst distances and Nei’s distances. The model with the greatest likelihood (logistic regression) was selected as the most likely model and divergence time (in generations) and population sizes were estimated from this model.

To complement analysis with DIY-ABC, TreeMix (Pickrell and Pritchard 2012) was used to investigate historical admixture events in eggplant without requiring prior hypotheses on the presence or absence of gene flow. Populations were grouped in the same way as for the DIYABC analysis. Treemix builds ML trees of the populations, and then models admixture events for any populations that show poor fit. 500 bootstrap replicates were conducted for each of the number of modelled migration events (from 1 to 10). Phylip *consense* (Felsenstein 2005) was used to produce a consensus tree from the bootstrap replicates and. BITE (Milanesi et al., 2017) was used to compute the fraction of variance explained by each migration model, the log likelihood, and the number of statistically significant migrations (p<0.05), and to visualise the consensus trees with bootstrap values and migration edges.

***Transcriptome Sequence Analysis***

Fastq formatted reads from the WTCHG were trimmed of adapters and low quality sequences (phred quality < 5), and short sequences (< 36 nucleotides) were removed using Trimmomatic v. 0.32 ([Bolger, et al. 2014](#_ENREF_9)) with settings LEADING:5 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:36. Only those reads which remained as part of a pair of reads were considered further.

The pairs of files from the 28 individuals were normalised with the script insilico\_read\_normalization.pl (removing reads with kmer coverage > 30) from within the Trinity RNAseq analysis package (ver. 2.0.6; ([Grabherr, et al. 2011](#_ENREF_38)). These were then assembled with Trinity with settings --min\_kmer\_cov 2, --max\_internal\_gap\_same\_path 15, and --max\_diffs\_same\_path 4. These settings were chosen to ignore single copy (potentially error-containing) kmers, and to allow two reads differing by up to a 15b indel and/or up to four SNPs (single nucleotide polymorphisms) to be assembled, primarily because multiple individuals were being assembled and indels and SNPs were predicted to be common. Trinity assembles the data into ’components’ (loosely akin to genes) and ‘transcripts’, where each component is made up of one or more transcripts, hence these primarily represent alternatively-spliced forms of the component ([Grabherr, et al. 2011](#_ENREF_38)). The terms components and transcripts will be used hereafter. *de novo* assembly was selected instead of mapping to the published eggplant genome ([Hirakawa, et al. 2014](#_ENREF_43)) because of the highly fragmented and incomplete nature of the genome. Reads from each sample were then mapped back to the reference transcriptome in *Trinity* using bowtie ([Langmead, et al. 2009](#_ENREF_52)) and RSEM ([Li and Dewey 2011](#_ENREF_54)) to give DNA sequence alignments and expression per transcript. To reduce the number of spurious transcripts (as evidenced by very low expression, or by a transcript corresponding to a low percentage of the sum of all transcripts for a component) the transcriptome was reduced to only those transcripts expressed at FPKM (Fragments Per Kilobase of transcript per Million mapped reads) ≥ 1 and expressed at ≥ 5% of the total expression for that component. Expression was analysed using edgeR ([Robinson, et al. 2010](#_ENREF_75)) and a significance threshold of FDR < 0.05.

Following transcriptome assembly and trimming, each transcript was compared to the tomato genome CDS database (ITAG ver. 2.4; ftp://ftp.solgenomics.net/tomato\_genome) and the Arabidopsis CDS database (TAIR ver. 10; https://www.arabidopsis.org) using BLAST ([Altschul, et al. 1997](#_ENREF_3)) and an e-value cut-off of e-30. To identify SNPs between the accessions, the bam files from the read mapping were parsed through SAMtools (ver 0.1.19; ([Li, et al. 2009](#_ENREF_55)) with mpileup settings -q 3 -Q 20 and bcftools (ver 1.2.1; https://github.com/samtools/bcftools) with setting -d 3. This converted the bam files into vcf files, and subsequently into fastq then fasta files (perl script available upon request).

For the population genetic analyses we trimmed the transcriptome to a single transcript per component (keeping the longest transcript). This reduces any potential bias that could result from some SNPs being included more than once (because the SNP is present in multiple transcripts from the same component). These transcriptome contigs were loaded into Proseq ([Filatov 2009](#_ENREF_33)) as a reference and then the 28 fasta files loaded into Proseq. To reduce the data to only sites present in all individuals we first removed short sequences (<100b) and then loci with less that a minimum number of individuals present in each group (see text).

Nucleotide diversity (π; [Nei and Li 1979](#_ENREF_68)) was calculated in Proseq for each group. Fst, Wright’s fixation index ([Wright 1951](#_ENREF_86)), and Dxy, Nei’s total nucleotide diversity ([Nei 1987](#_ENREF_67)), between subpopulations, was calculated for loci where ≥100 bp of sequence was present in the minimum number of individuals as stated above for both groups being compared. Dxy is not affected by intra-population diversity, whereas Fst is (reviewed in ([Cruickshank and Hahn 2014](#_ENREF_19)). Loci in the top 5% of the Fst and Dxy distribution were identified, with the expectation that this group being enriched for loci showing greater than expected divergence between the wild and landrace populations.

The vcf files were used to predict whether SNPs in coding regions were synonymous or non-synonymous. This was carried out for the individuals compared in the selection analysis (see below for justification) and the three outgroups (to polarise the SNPs). Open reading frames (ORFs) > 300 bp were identified from the reference transcriptome using getorf (http://emboss.sourceforge.net/apps/cvs/emboss/apps/getorf.html) and the longest from each contig was presumed to be the most likely. From the vcf files, multi-allelic SNPs, indels, and SNPs outside putative ORFs were ignored. SNP calls for individuals with less than 20 reads covering the SNP and a phred score of less than 10 were excluded. For the remaining SNPs the reference and alternative amino acid encoded at the site of the SNP were identified and recorded for each individual. SNPs present at a frequency <20% in one population and >80% in the other were deemed to be ‘strongly differentiated’. Of these, non-synonymous SNPs were given a Grantham score ([Grantham 1974](#_ENREF_39)) and category (conservative, moderately conservative, moderately severe, severe) based on severity of amino acid replacement.

**Gene Ontology (GO) analysis of domestication**

To determine if particular pathways or functions of genes appeared to be under selection during eggplant domestication, we used agriGO ([Du, et al. 2010](#_ENREF_24)) to determine if there were any Gene Ontology (GO) terms ([Ashburner, et al. 2000](#_ENREF_6)) significantly over-represented in the lists of candidate genes produced from the different analyses of selection. The best *Arabidopsis* hit for each gene in each list was identified based on comparison to the *Arabidopsis* peptide sequences (TAIR ver. 10; blastx; e cut-off of e-30).

**ACKNOWLEDGMENTS**

This work was supported by an “Annual Adventures in Research” grant from the University of Southampton to M.A.C. and J.G. and a studentship from the University of Southampton for A.M.L.P. We would like to thank the staff of the Iridis HPC Facility at the University of Southampton, especially Elena Vataga, for their computational support throughout. We also thank the Genomic Diversity Facility at Cornell University and the Wellcome Trust Centre for Human Genetics for carrying out the GBS and transcriptome sequencing, respectively. We are grateful to members of the Chapman lab and other colleagues at the University of Southampton for comments on earlier versions of this manuscript.

**REFERENCES**

Aguilar-Martinez JA, Sinha N 2013. Analysis of the role of Arabidopsis class I TCP genes AtTCP7, AtTCP8, AtTCP22, and AtTCP23 in leaf development. Frontiers in Plant Science 4: 406. doi: 10.3389/fpls.2013.00406

Allaby RG, Fuller DQ, Brown TA 2008. The genetic expectations of a protracted model for the origins of domesticated crops. Proceedings of the National Academy of Sciences of the United States of America 105: 13982-13986.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.

Arias DM, Rieseberg LH 1994. Gene flow between cultivated and wild sunflowers. Theoretical and Applied Genetics 89: 655-660.

Arnold ML 2004. Natural hybridization and the evolution of domesticated, pest and disease organisms. Molecular Ecology 13: 997-1007.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29.

Bellucci E, Bitocchi E, Ferrarini A, Benazzo A, Biagetti E, Klie S, Minio A, Rau D, Rodriguez M, Panziera A, Venturini L, Attene G, Albertini E, Jackson SA, Nanni L, Fernie AR, Nikoloski Z, Bertorelle G, Delledonne M, Papa R 2014. Decreased nucleotide and expression diversity and modified coexpression patterns characterize domestication in the common bean. Plant Cell 26: 1901-1912. doi: 10.1105/tpc.114.124040

Bitocchi E, Bellucci E, Giardini A, Rau D, Rodriguez M, Biagetti E, Santilocchi R, Zeuli PS, Gioia T, Logozzo G, Attene G, Nanni L, Papa R 2013. Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. New Phytologist 197: 300-313. doi: 10.1111/j.1469-8137.2012.04377.x

Bolger A, Lohse M, Usadel B 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Burke JM, Burger JC, Chapman MA 2007. Crop evolution: from genetics to genomics. Current Opinion in Genetics & Development 17: 525-532.

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, Polato NR, Olsen KM, Nielsen R, McCouch SR, Bustamante CD, Purugganan MD 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genetics 3: 1745-1756.

Carmel-Goren L, Liu YS, Lifschitz E, Zamir D 2003. The *SELF-PRUNING* gene family in tomato. Plant Molecular Biology 52: 1215-1222. doi: 10.1023/b:plan.0000004333.96451.11

Casa AM, Mitchell SE, Hamblin MT, Sun H, Bowers JE, Paterson AH, Aquadro CF, Kresovich S 2005. Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. Theoretical and Applied Genetics 111: 23-30.

Chakrabarti M, Zhang N, Sauvage C, Munos S, Blanca J, Canizares J, Jose Diez M, Schneider R, Mazourek M, McClead J, Causse M, van der Knaap E 2013. A cytochrome P450 regulates a domestication trait in cultivated tomato. Proceedings of the National Academy of Sciences of the United States of America 110: 17125-17130. doi: 10.1073/pnas.1307313110

Chapman MA, Hiscock SJ, Filatov DA 2016. The genomic bases of morphological divergence and reproductive isolation driven by ecological speciation in *Senecio* (Asteraceae). Journal of Evolutionary Biology 29: 98-113.

Chapman MA, Hiscock SJ, Filatov DA 2013. Genomic divergence during speciation driven by adaptation to altitude. Molecular Biology and Evolution 30: 2553-2567. doi: 10.1093/molbev/mst168

Chapman MA, Pashley CH, Wenzler J, Hvala J, Tang S, Knapp SJ, Burke JM 2008. A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). The Plant Cell 20: 2931-2945.

Chifman J, Kubato L 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30: 3317-3324

Cornuet J-M, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R, Marin J-M, Estoup A 2014. DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. Bioinformatics 30: 1187-1189. doi: 10.1093/bioinformatics/btt763

Cruickshank TE, Hahn MW 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular Ecology 23: 3133-3157. doi: 10.1111/mec.12796

Darwin CR. 1868. The Variation of Animals and Plants under Domestication: John Murray, London.

Daunay M-C, Lester RN, Ano G. 2001. Eggplant. In: Charrier A, Jacquot M, Hamon S, Nicolas D, editors. Tropical Plant Breeding. Montpellier: Science Publishers: Montpellier: Science Publishers. p. 199-222.

de Wet JMJ, Harlan, JR. 1975. Weeds and domesticates: Evolution in the man-made habitat. Economic Botany 29: 99-107.

Doganlar S, Frary A, Daunay MC, Lester RN, Tanksley SD 2002. Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant. Genetics 161: 1713-1726.

Doyle JJ, Doyle JL 1990. Isolation of plant DNA from fresh tissue. Focus 12: 13-15.

Du Z, Zhou X, Ling Y, Zhang Z, Su Z 2010. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Research 38: W64–W70.

Earl DA, von Holdt BM 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4: 359-361.

Eaton DA 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30: 1844-1849. doi: 10.1093/bioinformatics/btu121

Elitzur T, Nahum H, Borovsky Y, Pekker I, Eshed Y, Paran I 2009. Co-ordinated regulation of flowering time, plant architecture and growth by *FASCICULATE*: the pepper orthologue of *SELF PRUNING*. Journal of Experimental Botany 60: 869-880. doi: 10.1093/jxb/ern334

Ellstrand NC. 2003. Dangerous liaisons? When cultivated plants mate with their wild relatives. Baltimore: The Johns Hopkins University Press.

Ellstrand NC, Heredia SM, Leak-Garcia JA, Heraty JM, Burger JC, Yao L, Nohzadeh-Malakshah S, Ridley CE 2010. Crops gone wild: evolution of weeds and invasives from domesticated ancestors. Evolutionary Applications 3: 494-504. doi: 10.1111/j.1752-4571.2010.00140.x

Ellstrand NC, Meirmans P, Rong J, Bartsch D, Ghosh A, Jong TJd, Haccou P, Lu B-R, Snow AA, Jr. CNS, Strasburg JL, Tienderen PHv, Vrieling K, Hooftman D 2013. Introgression of crop alleles into wild or weedy populations. Annual Review of Ecology, Evolution, and Systematics 44: 325-345. doi: 10.1146/annurev-ecolsys-110512-135840

Ellstrand NC, Prentice HC, Hancock JF 1999. Gene flow and introgression from domesticated plants into their wild relatives. Annual Review of Ecology and Systematics 30: 539-563.

Falush D, Stephens M, Pritchard JK 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.

Felsenstein J 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Filatov DA 2009. Processing and population genetic analysis of multigenic datasets with ProSeq3 software. Bioinformatics 25: 3189-3190. doi: 10.1093/bioinformatics/btp572

Frary A, Frary A, Daunay M-C, Huvenaars K, Mank R, Doğanlar S 2014. QTL hotspots in eggplant (*Solanum melongena*) detected with a high resolution map and CIM analysis. Euphytica 197: 211-228. doi: 10.1007/s10681-013-1060-6

Frary A, Nesbitt TC, Frary A, Grandillo S, van der Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB, Tanksley SD 2000. *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. Science 289: 85-88.

Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, Pretty J, Robinson S, Thomas SM, Toulmin C 2010. Food security: The challenge of feeding 9 billion people. Science 327: 812-818. doi: 10.1126/science.1185383

Goldsbrough A, Belzile F, Yoder JI 1994. Complementation of the tomato *anthocyanin without* (*aw*) nutant using the dihydroflavonol 4-reductase gene. Plant Physiology 105: 491-496. doi: 10.1104/pp.105.2.491

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD, Chen ZH, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29: 644-U130. doi: 10.1038/nbt.1883

Grantham R 1974. Amino acid difference formula to help explain protein evolution. Science 185: 862-864.

Gressel J. 2005. Crop Ferality and Volunteerism: CRC Press, Boca Raton, Florida.

Harlan JR. 1992. Crops and man. Madison, WI, USA: American Society for Agronomy.

Hay FR, Probert RJ 2013. Advances in seed conservation of wild plant species: a review of recent research. Conservation Physiology 1: cot030. doi: 10.1093/conphys/cot030

Hirakawa H, Shirasawa K, Miyatake K, Nunome T, Negoro S, Ohyama A, Yamaguchi H, Sato S, Isobe S, Tabata S, Fukuoka H 2014. Draft genome sequence of eggplant (*Solanum melongena* L.): the representative *Solanum* species indigenous to the Old World. DNA Research 21: 649-660. doi: 10.1093/dnares/dsu027

Honaas LA, Wafula EK, Wickett NJ, Der JP, Zhang Y, Edger PP, Altman NS, Pires JC, Leebens-Mack JH, dePamphilis CW 2016. Selecting superior *de novo* transcriptome assemblies: Lessons learned by leveraging the best plant genome. PLoS One 11: e0146062. doi: 10.1371/journal.pone.0146062

Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J 2013. The genomic signature of crop-wild introgression in maize. PLoS Genetics 9: e1003477. doi: 10.1371/journal.pgen.1003477

Jakobsson M, Rosenberg NA 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801-1806. doi: 10.1093/bioinformatics/btm233

Janzen GM, Wang L, Hufford MB 2019. The extent of adaptive wild introgression in crops. New Phytologist 221: 1279-1288. doi:10.1111/nph.15457

Jayaraman K 2010. Bt brinjal splits Indian cabinet. Nature Biotechnology 28: 296-296.

Kashyap A, Weber S. 2013. Starch grain analysis and experiments provide insights into Harappan cooking practices. In: Abraham SA, Gullapalli P, Raczek TP, Rizvi UZ, editors. Connections and Complexity: New Approaches to the Archaeology of South Asia: Left Coast Press, Taylor and Francis, Oxford, UK. p. 177-193.

Knapp S, Vorontsova MS, 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. doi: 10.1371/journal.pone.0057039

Koenig D, Jimenez-Gomez JM, Kimura S, Fulop D, Chitwood DH, Headland LR, Kumar R, Covington MF, Devisetty UK, Tat AV, Tohge T, Bolger A, Schneeberger K, Ossowski S, Lanz C, Xiong G, Taylor-Teeples M, Brady SM, Pauly M, Weigel D, Usadel B, Fernie AR, Peng J, Sinha NR, Maloof JN 2013. Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. Proceedings of the National Academy of Sciences of the United States of America 110: E2655-E2662. doi: 10.1073/pnas.1309606110

Langmead B, Trapnell C, Pop M, Salzberg SL 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10: 1-10. doi: 10.1186/gb-2009-10-3-r25

Lester RN, Hasan SMZ. 1991. Origin and domestication of the brinjal eggplant, *Solanum melongena*, from *S. incanum* in Africa and Asia. In: Hawkes JG, Lester RN, Nee M, Estrada N, editors. Solanaceae III: Taxonomy, Chemistry, Evolution: Royal Botanic Gardens, Kew, UK. p. 369-387.

Li B, Dewey C 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079. doi: 10.1093/bioinformatics/btp352

Li L-F, Li Y-L, Jia Y, Caicedo AL, Olsen KM 2017. Signatures of adaptation in the weedy rice genome. Nature Genetics 49: 811-814. doi: 10.1038/ng.3825

Lin Z, Li X, Shannon LM, Yeh C-T, Wang ML, Bai G, Peng Z, Li J, Trick HN, Clemente TE, Doebley J, Schnable PS, Tuinstra MR, Tesso TT, White F, Yu J 2012. Parallel domestication of the *Shattering1* genes in cereals. Nature Genetics 44: 720-724. doi: 10.1038/ng.2281

Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, Campbell BC, Hu W, Innes DJ, Han X, Cruickshank A, Dai C, Frère C, Zhang H, Hunt CH, Wang X, Shatte T, Wang M, Su Z, Li J, Lin X, Godwin ID, Jordan DR, Wang J 2013. Whole-genome sequencing reveals untapped genetic potential in Africa’s indigenous cereal crop sorghum. Nature Communications 4: 2320. doi: 10.1038/ncomms3320

Mandel JR, Ramsey AJ, Iorizzo M, Simon PW 2016. Patterns of gene flow between crop and wild carrot, *Daucus carota* (Apiaceae) in the United States. PLoS One 11: e0161971. doi: 10.1371/journal.pone.0161971

McCouch SR, Baute GJ, Bradeen J, Bramel P, Bretting PK, Buckler E, Burke JM, Charest D, Cloutier S, Cole G, Dempewolf H, Dingkuhn M, Feuillet C, Gepts P, Grattapaglia D, Guarino L, Jackson S, Knapp S, Langridge P, Lawton-Rauh A, Lijua Q, Lusty C, Michael T, Myles S, Naito K, Nelson RL, Pontarollo R, Richards CM, Rieseberg LH, Ross-Ibarra J, Rounsley S, Sackville Hamilton R, Schurr U, Stein N, Tomooka, van der Knaap E, van Tassel D, Toll J, Valls J, Varshney RK, Ward J, Waugh R, Wenzl P, Zamir D 2013. Agriculture: feeding the future. Nature 499: 23-24.

Meyer RS, Karol KG, Little DP, Nee MH, Litt A 2012. Phylogeographic relationships among Asian eggplants and new perspectives on eggplant domestication. Molecular Phylogenetics and Evolution 63: 685-701.

Meyer RS, Purugganan MD 2013. Evolution of crop species: genetics of domestication and diversification. Nature Reviews Genetics 14: 840-852.

Meyer RS, Whitaker BD, Little DP, Wu S-B, Kennelly EJ, Long C-L, Litt A 2015. Parallel reductions in phenolic constituents resulting from the domestication of eggplant. Phytochemistry 115: 194-206. doi: https://doi.org/10.1016/j.phytochem.2015.02.006

Milanesi M, Capomaccio S, Vajana E, Bomba L, Garcia JF, Ajmone-Marsan P, Colli L. 2017. BITE: an R package for biodiversity analyses. bioRxiv. https://doi.org/10.1101/181610

Mishina TE, Zeier J 2007. Pathogen-associated molecular pattern recognition rather than development of tissue necrosis contributes to bacterial induction of systemic acquired resistance in Arabidopsis. Plant Journal 50: 500-513. doi: 10.1111/j.1365-313X.2007.03067.x

Miya A, Albert P, Shinya T, Desaki Y, Ichimura K, Shirasu K, Narusaka Y, Kawakami N, Kaku H, Shibuya N 2007. CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in *Arabidopsis*. Proceedings of the National Academy of Sciences of the United States of America 104: 19613-19618. doi: 10.1073/pnas.0705147104

Mutegi E, Snow AA, Rajkumar M, Pasquet R, Ponniah H, Daunay M-C, Davidar P 2015. Genetic diversity and population structure of wild/weedy eggplant (*Solanum insanum*, Solanaceae) in southern India: Implications for conservation. American Journal of Botany 102: 140-148. doi: 10.3732/ajb.1400403

Nei M. 1987. Molecular Evolutionary Genetics: Columbia University Press, New York, NY.

Nei M, Li WH 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of the United States of America 76: 5269-5273.

Niyogi KK, Grossman AR, Björkman O 1998. Arabidopsis mutants define a central role for the xanthophyll cycle in the regulation of photosynthetic energy conversion. The Plant Cell 10: 1121-1134. doi: 10.1105/tpc.10.7.1121

Olsen KM, Purugganan MD 2002. Molecular evidence on the origin and evolution of glutinous rice. Genetics 162: 941-950.

Ori N, Eshed Y, Paran I, Presting G, Aviv D, Tanksley S, Zamir D, Fluhr R 1997. The *I2C* family from the wilt disease resistance locus *I2* belongs to the nucleotide binding, leucine-rich repeat superfamily of plant resistance genes. The Plant Cell 9: 521-532. doi: 10.1105/tpc.9.4.521

Pickrell JK, Pritchard JK 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics, 8: e1002967.

Poppy GM, Chiotha S, Eigenbrod F, Harvey CA, Honzak M, Hudson MD, Jarvis A, Madise NJ, Schreckenberg K, Shackleton CM, Villa F, Dawson TP 2014. Food security in a perfect storm: using the ecosystem services framework to increase understanding. Philosophical Transactions of the Royal Society B-Biological Sciences 369. doi: 10.1098/rstb.2012.0288

Portis E, Barchi L, Toppino L, Lanteri S, Acciarri N, Felicioni N, Fusari F, Barbierato V, Cericola F, Valè G, Rotino GL 2014. QTL mapping in eggplant reveals clusters of yield-related loci and orthology with the tomato genome. PLoS One 9: e89499. doi: 10.1371/journal.pone.0089499

Purugganan MD, Fuller DQ 2009. The nature of selection during plant domestication. Nature 457: 843-848. doi: 10.1038/nature07895

Robinson MD, McCarthy DJ, Smyth GK 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140.

Rodriguez GR, Munos S, Anderson C, Sim S-C, Michel A, Causse M, Gardener BBM, Francis D, van der Knaap E 2011. Distribution of *SUN*, *OVATE*, *LC*, and *FAS* in the tomato germplasm and the relationship to fruit shape diversity. Plant Physiology 156: 275-285. doi: 10.1104/pp.110.167577

Ross-Ibarra J, Morrell PL, Gaut BS 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proceedings of the National Academy of Sciences of the United States of America 104: 8641-8648.

Saito K, Kobayashi M, Gong Z, Tanaka Y, Yamazaki M 1999. Direct evidence for anthocyanidin synthase as a 2-oxoglutarate-dependent oxygenase: molecular cloning and functional expression of cDNA from a red forma of *Perilla frutescens*. Plant Journal 17: 181-189.

Swanson-Wagner R, Briskine R, Schaefer R, Hufford MB, Ross-Ibarra J, Myers CL, Tiffin P, Springer NM 2012. Reshaping of the maize transcriptome by domestication. Proceedings of the National Academy of Sciences of the United States of America 109: 11878-11883. doi: 10.1073/pnas.1201961109

Tanksley SD, McCouch SR 1997. Seed banks and molecular maps: Unlocking genetic potential from the wild. Science 277: 1063-1066.

Verburg JG, Huynh QK 1991. Purification and characterization of an antifungal chitinase from *Arabidopsis thaliana*. Plant Physiology 95: 450-455. doi: 10.1104/pp.95.2.450

Vernot B, Akey JM 2014. Resurrecting surviving neandertal lineages from modern human genomes. Science 343: 1017-1021. doi: 10.1126/science.1245938

Wang H, Vieira FG, Crawford JE, Chu C, Nielsen R 2017. Asian wild rice is a hybrid swarm with extensive gene flow and feralization from domesticated rice. Genome Research 27: 1029–1038. doi: 10.1101/gr.204800.116

Wang JX, Gao TG, Knapp S 2008. Ancient Chinese literature reveals pathways of eggplant domestication. Annals of Botany 102: 891-897. doi: 10.1093/aob/mcn179

Weese TL, Bohs L 2010. Eggplant origins: Out of Africa, into the Orient. Taxon 59: 49-56.

Wright S 1951. The genetical structure of populations. Annals of Eugenics 15: 323-354.

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS 2005. The effects of artificial selection on the maize genome. Science 308: 1310-1314.

Yu JW, Rubio V, Lee NY, Bai S, Lee SY, Kim SS, Liu L, Zhang Y, Irigoyen ML, Sullivan JA, Zhang Y, Lee I, Xie Q, Paek NC, Deng XW 2008. COP1 and ELF3 control circadian function and photoperiodic flowering by regulating GI stability. Molecular Cell 32: 617-630. doi: 10.1016/j.molcel.2008.09.026

Zeilmaker T, Ludwig NR, Elberse J, Seidl MF, Berke L, Van Doorn A, Schuurink RC, Snel B, Van den Ackerveken G 2015. DOWNY MILDEW RESISTANT 6 and DMR6-LIKE OXYGENASE 1 are partially redundant but distinct suppressors of immunity in Arabidopsis. The Plant Journal 81: 210-222. doi: doi:10.1111/tpj.12719

Zizumbo-Villarreal D, Colunga-GarciaMarin P, Payro de la Cruz E, Delgado-Valerio P, Gepts P 2005. Population structure and evolutionary dynamics of wild-weedy-domesticated complexes of common bean in a Mesoamerican region. Crop Science 45: 1073-1083.

**FIGURE LEGENDS**

**Figure 1 – Relationships between wild and domesticated eggplant.** (A) Maximum likelihood phylogenetic tree of wild and domesticated eggplants based on 4054 single nucleotide polymorphisms. Bootstrap values greater than 70 are indicated and nodes are coloured according to the legend. Arrows indicate individuals for which transcriptome sequences were obtained (see text for details). Inset depicts the ‘gradation’ between wild and domesticated eggplants in detail. (B) Output from STRUCTURE run with K = 2 representing proportional assignment of each individual to two clusters. (C) Output from STRUCTURE run with K = 5 representing proportional assignment of each individual to five clusters. A tree containing accessions names for each individual is presented in supplementary fig. S1. (D and E) DIYABC analysis of population divergence scenarios. (D) represents the analysis of two groups of accessions in the ‘gradation’ between wild and domesticated eggplants (green boxes in inset of 1A); (E) represents the analysis of the two pairs of ‘wild-like’ accessions nested within the domesticates. In both, the most supported scenario is shown surrounded by a box and the parameter estimates are given in the table below. tx represents coalescence time (in generations), Nx represents estimated population size, and rx represents the proportion of gene flow.

**Figure 2 – Boxplots depicting (A) transcriptome sequence polymorphism within each of the four groups/species investigated; (B) genetic differentiation (Dxy) between populations; and (C) genetic differentiation (Fst) between populations.** Lower ends, and in some cases medians, of the distributions were 0 in some cases (i.e. for OVI and MEL in panel A, and INS-OVI and OVI-MEL in panel C).

**Figure 3 – Boxplots depicting genetic differentiation between populations separated into differentially expressed (DE) loci (grey), and non-DE loci (white).** (A) Fst, (B) Dxy. In all cases DE loci exhibited significantly higher differentiation than the non-DE loci. Lower ends, and in some cases medians, of the distributions were 0 in some cases (DE loci in the Fst panel).

**Figure S1 – Relationships between wild and domesticated eggplant**. A.This is the same maximum likelihood phylogenetic tree as in Fig. 1 but with full accession names. B. SVDquartets analysis using the same data. The position of the putative feral eggplants are highlighted in both A and B to demonstrate their inconsistent phylogenetic positions.

**Figure S2 – DeltaK plot for the different *K* values in STRUCTURE presented in Figure 1.**

**Figure S3 – Results of the TreeMix analysis.** (A) Statistics associated with the models examining 1-10 migrations, (B) variance explained (f) for each model, (C) Output tree for the 3 migration model. Lines between groups represent migrations, coloured by weight.

**Figure S4 – Representative photographs of wild (A) and cultivated (B) eggplant leaves.** (A) accession MM1545 from Malaysia (inset shows the purple pigmentation at the base of the spine); (B) accession PI470273 from Indonesia. Scale bar is in 10 mm gradations.

**Figure S5 – Classification of SNPs, according to severity of amino acid substitution,** for (A) the high Dxy loci vs the rest of the loci, and (B) SNPs segregating in the wild and landrace populations.

Figure 1

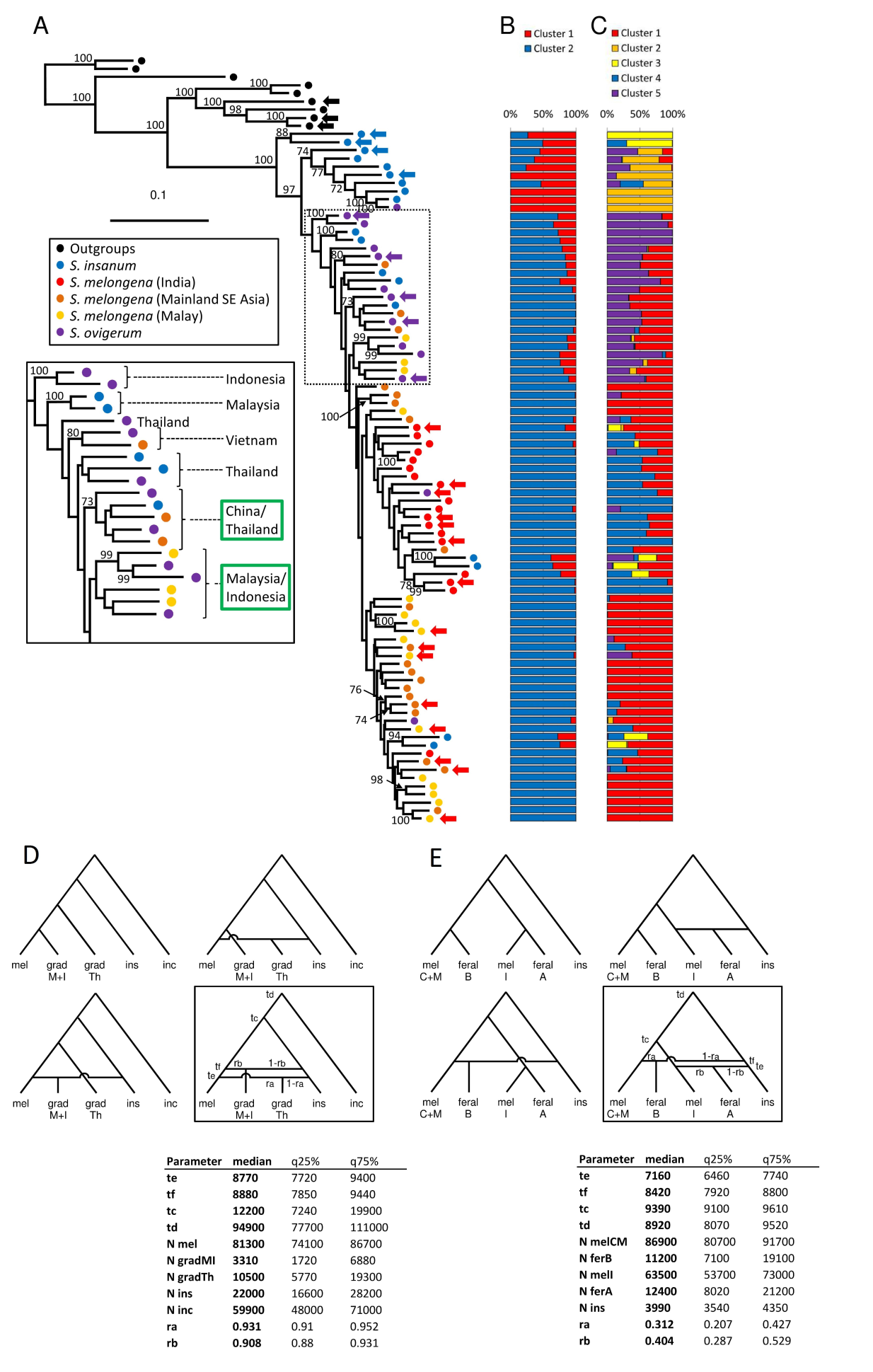


Figure 2

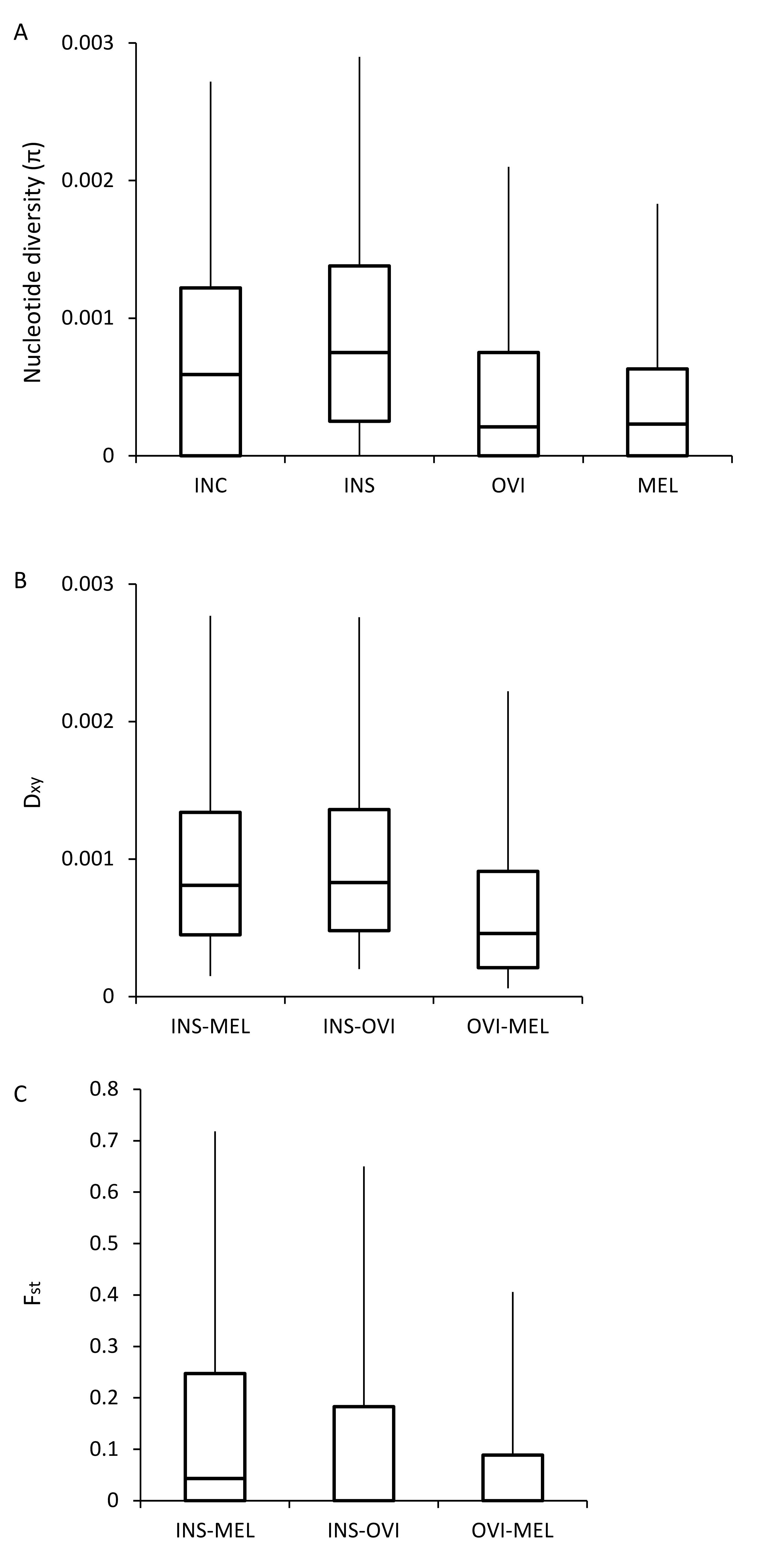


Figure 3

