Identifying genomic regions targeted during eggplant domestication using transcriptome data

Anna M. L. Page[1] and Mark A. Chapman[1,*]

[1] Biological Sciences, University of Southampton, Southampton, SO17 1BJ, UK.

* Author for correspondence: m.chapman@soton.ac.uk; +44(0)2380 594396

ABSTRACT

Identifying genes and traits that have diverged during domestication provides key information of importance for maintaining and even increasing yield and nutrients in existing crops. A 'bottom up' population genetics approach was used to identify signatures of selection across the eggplant genome, to better understand the process of domestication. RNA-seq data was obtained for four wild eggplants (*Solanum insanum* L.) and 16 domesticated eggplants (*S. melongena* L.) and mapped to the eggplant genome. SNPs exhibiting signatures of selection in domesticates were identified as those exhibiting high F_{ST} between the two populations (evidence of significant divergence) and low π for the domesticated population (indicative of a selective sweep). Some of these regions appear to overlap with previously identified QTL for domestication traits. Genes in regions of linkage disequilibrium surrounding these SNPs were searched against the *Arabidopsis thaliana* and tomato genomes to find orthologues. Subsequent Gene Ontology (GO) enrichment analysis identified over-representation of GO terms related to photosynthesis and response to the environment. This work reveals genomic changes involved in eggplant domestication and improvement, and how this compares to observed changes in the tomato genome, revealing shared chromosomal regions involved in the domestication of both species.

INTRODUCTION

By studying domestication, the human population can better understand events of the past and arm themselves against an uncertain future by breeding crops that can meet the demands of climate change and a growing population (Godfray et al., 2010). Despite being the second most important solanaceous fruit crop after tomato (Knapp et al., 2013), eggplant is relatively understudied. Recent advancements, such as the development of a reference genome and backcross populations for breeding, make this an ideal time to study the domestication of eggplant (Chapman, 2019).

During the process of domestication, eggplants have been exposed to strong, directional, human-mediated selection. This established a suite of domestication syndrome traits that distinguish domestic eggplants from the wild progenitor *Solanum insanum* L., for example a loss of spines, larger and non-bitter fruits, and more consistent flowering (Page et al., 2019b). However, because landraces and wild relatives typically occupy the same range, gene flow and hybridisation are common between wild and domesticated populations (Davidar et al., 2015, Page et al., 2019a). The complicated demographics this produces can make it challenging to identify true wild and domesticated accessions based on phenotype alone. For example, Knapp et al. (2013) proposed a key to morphological characteristics for distinguishing ambiguous specimens; however, we recently showed that some phenotypically wild eggplants were in fact feral domesticates with admixed (wild and domesticated) genomes (Page et al. 2019a). Analyses in eggplant aiming to identify variants associated with selection during domestication must therefore be accompanied by demographic inference such as a phylogeny to verify true wild and true domesticated accessions.

Previous work to identify the genomic regions underlying phenotypes in eggplants have largely taken a top-down approach, most often analyses of quantitative trait loci (QTL). These analyses have revealed loci linked to domestication traits using crosses between domestic and wild eggplant parents (Doganlar et al., 2002, Frary et al., 2014, Miyatake et al., 2020, Portis et al., 2014, Toppino et al., 2016), and loci controlling specific traits that vary between cultivars such as disease resistance (Barchi et al., 2018), and anthocyanin content (Barchi et al., 2012).

Domestication imposes a bottleneck on the crop, which increases the strength of genetic drift. This will lead to a genome-wide reduction in diversity. Regions of the genome which have been under selection will have lower diversity still, driving the selected allele and the genomic regions surrounding it towards fixation (Burke et al. 2007; Olsen and Wendel, 2013). The eggplant reference genome recently developed by Barchi et al. (2019) brings the opportunity to identify regions under selection using a bottom-up approach; therefore we used this to scan the genome for signals of selection without making *a priori* choices about the phenotypes of interest (Ross-Ibarra et al., 2007).

Previously we analysed RNA-seq data on a gene-by-gene basis (Page et al. 2019a) as there was no reference genome available to us. This means that outlier loci we identified based on patterns of nucleotide diversity could be influenced by loci in linkage disequilibrium (LD). We therefore chose to reanalyse this data mapping the RNA-seq reads to the genome to identify more clearly patterns of selection during domestication. Looking at overall patterns of selection on the genome, and how these compare to other Solanaceous crops also reveals more about the effects artificial selection has had on the genome, such as what proportion of the genome was under selection, and whether the same regions were targeted in related crops.

MATERIALS AND METHODS

Existing Datasets, RNA extraction and sequencing

The RNA-seq data from Page et al. (2019a) was reanalysed in this work. We used the phylogenetic analysis of a genotyping-by-sequencing (GBS) dataset of 95 domesticated eggplants and wild relative species sampled across their range (Page et al. 2019a) to define true wild and domesticated eggplants. True wilds are defined as wild accessions sister to the domesticates in the phylogeny, while phenotypically wild accessions nested within the domesticated clade with admixed genomes are defined as feral and are excluded here.

Briefly, one fully expanded leaf from each accession (four wild and 16 domesticated; Table 1) was frozen in liquid nitrogen and RNA was extracted using a Qiagen RNeasy Plant Mini kit (Qiagen, UK), utilising an on-column DNase step (Qiagen). Samples were sent to the Wellcome Trust Centre for Human Genetics (WTCHG, Oxford, UK) for quantification, quality checking and subsequent library preparation using Illumina's Stranded Truseq kit (Illumina, UK). Up to 12 libraries (individually barcoded) were sequenced per lane on Illumina Hiseq2000 for 101 cycles (paired end). Samples were then de-multiplexed and bioinformatic analyses took on the University of Southampton Iridis4 supercomputer. All samples are available from the NCBI Sequence Read Archive (PRJNA526115).

Transcriptome variant calling

RNA-seq reads were trimmed and poor quality bases and short reads removed using Trimmomatic v. 0.32 (Bolger et al., 2014) as described previously (Page et al., 2019a), and aligned to the eggplant genome (Barchi et al., 2019) using STAR (Dobin et al., 2012). A gff3 file of predicted genes from the eggplant genome was supplied when indexing the genome with STAR, which improved mapping quality by providing annotations for known splice junctions. After mapping, Picard 2.8.3 (Broad Institute, 2019) and GATK (McKenna et al., 2010) were used to process the mapped reads following the recommendations of GATK best practices for calling variants in RNA-seq, i.e. marking duplicates in Picard, using SplitNCigarReads in GATK to split reads into exon segments, and calling variants with HaplotypeCaller in GATK. Hard filtering was then applied to the callset. GATK best practices recommend filtering clusters of 3 or more SNPs within a 35-base window, Fisher Strand values > 30 and Quality by Depth values < 2. BCFtools within SAMtools v. 0.1.19 (Li et al., 2009) was used to merge the vcf files of samples into two files (wild and domesticated), then further filtering was done to remove any SNPs with a MAF of < 0.1, and with coverage in the lowest 99th percentile.

Population genomic statistics

The *Populations* program in Stacks v. 1.48 (Catchen et al., 2013), with the vcf files as input, was used to calculate AMOVA F_{ST} between the domesticated and wild populations, and π within the domesticated population. SNPs were not filtered for missing data because of the small sample size in the wild population, which we assume adds some noise to the estimates of F_{ST} and π. AMOVA F_{ST} in Stacks is derived from Weir (1996), and was calculated as the weighted average of the surrounding 450kb (the default value in Stacks) of sequence for each SNP. A weighted average of π in the

surrounding 450kb of sequence was also calculated for the domesticated population. 10^6 bootstrap replicates were performed and used to report *P* values, which were subsequently corrected using the Holm-Bonferroni sequential correction (Holm, 1979), calculated using the p.adjust command in *R* 3.1.3 (Team, 2015) on a per chromosome basis. These were considered outlier SNPs for high F_{ST}, low π and their overlap.

To identify the region of the genome in linkage disequilibrium (LD) with outlier loci, haplotype blocks containing SNPs in "strong LD" (*sensu* PLINK) were estimated from the filtered combined wild and domesticated dataset using PLINK 1.9 (Purcell et al., 2007, Chang et al., 2015). The cumulative percentage of haplotype blocks was then plotted against size of haplotype blocks and used to identify a point at which the curve levelled off. This was used to define regions of interest around outlier loci, and from within these regions, genes were identified.

Gene Ontology analysis

A fasta file containing the sequences of all annotated genes in the regions putatively under selection was created from the eggplant genome sequence using gff3 files to identify genomic positions. This fasta file was used in a BLAST search against *Arabidopsis thaliana* CDS sequences (ver. TAIR10) and tomato (ver. ITAG2.4) using Bioedit (Hall, 1999), retaining significant hits (e-value < e^{-10}). These lists of genes were then used in a gene ontology (GO) enrichment analysis (Ashburner et al., 2000) with agriGo (Du et al., 2010) to identify GO terms that were over-represented in the candidate gene list. False discovery rate correction was applied to the P values.

RESULTS

Mapping and calling SNPs

Between 11.7 and 26.9 M reads were sequenced from each accession, with an average of 19.9M (± 0.9 M [SE]). Annotated coding genes make up 8.90% of the 1.2 Gb eggplant genome and of the 27,139 annotated coding genes in the eggplant reference genome, 22,954 were present in the transcriptome of at least one accession. Following hard filtering on the SNPs called, the average number of SNPs per sample decreased by 10.73%, from 117,699 to 105,075. After merging samples based on population, further filtering on MAF and coverage retained 203,611 out of a possible 668,552 SNPs.

Haplotype block analysis to estimate the average LD in the eggplant genome

When the upper-class boundary of the estimated haplotype block sizes, estimated for the combined wild and domesticated data, was plotted against the cumulative percentage, there was a clear drop off in the rate of increase of haplotype block size (Figure 1). The majority of haplotype blocks (72.5%) are under 3.93 Mb in size, and following this point, increasing the length of block has diminishing returns for the increase in the cumulative percentage of haplotype blocks. Therefore 3.93 Mb was used to define the region of interest around loci with significant F_{ST} and/or π.

Regions of the genome under selection

After combining significant SNPs within 3.93 MB as potentially in LD and therefore marking the same region of the genome, six genomic regions with significantly (adjusted $P < 0.05$) low π in domesticates were identified (on chromosomes [chr] 1, 2 [2 regions], 4, 6 and 9), and three with significantly (adjusted $P < 0.05$) high F_{ST} were detected (on chr 2, 6 and 9; Table 2; Figure 2). The low π and high F_{ST} regions on chr 9 overlapped almost completely, whereas the others did not. There are other peaks in F_{ST}, such as at the ends of chromosomes 1, 3 and 8, and a highly heterogeneous distribution of π but these are not statistically significant. These eight regions totalled ca. 63 MB, approximately 5.6% of the eggplant genome and contained between 39 and 526 (mean 178.75) annotated genes. Of these, 375 had a significant (e value < e^{-10}) BLAST hit in Arabidopsis and 1113 had a significant BLAST hit in tomato (Supplementary Table 1). None of the putative orthologues had clear functions related to domestication traits in eggplant. In addition, we identified the genomic coordinates of molecular markers under the QTL peaks from the Doganlar et al. (2002) study and the prickliness indel marker from (Miyatake et al., 2020) (Supplementary Table 2). In only one case was there overlap with a selective sweep; this was for the shoot anthocyanin QTL *sa6.1* which overlapped with sweep 6_2 on chr 6.

GO analysis reveals putative selection on the photosynthetic pathway

A GO analysis using the combined list genes in the eight genomic regions putatively under selection was carried out, revealing significant ($P < 0.05$ after FDR correction) over-representation for 171 GO terms (Supplementary Table 3). Many of the over-represented GO terms are related to photosynthesis, including *photosynthesis*, *photosystems I and II*, *chlorophyll* and *chloroplasts*, *thylakoids*, *plastids,* and *tetrapyrrole binding*. This included the putative orthologues of genes *PsbO-1* (*PS II oxygen-evolving complex*) and *LHCB5* (*light harvesting complex of photosystem II 5*), both of which are involved in photosystem II assembly (Popelkova and Yocumab, 2011; Jansson 1999).

Other terms involved in response to the environment including *response to osmotic stress* and *response to salt stress* as well as *response to hormone stimulus* were also uncovered.

DISCUSSION

In this study, measures of population differentiation between wild and domesticated populations of eggplants and nucleotide diversity within the domesticated population were used to identify genomic regions having the signature of selection during domestication.

Using our pipeline, identifying haplotype blocks in which outlier SNPs were found, we find that genomic regions putatively under selection during eggplant domestication were few in number and are each relatively small in size, ranging between 7.8 and 8.3 Mb and are found on five of the eggplant chromosomes (chromosomes 1, 2, 4, 6 and 9). In total this comprises about 5.6% of the eggplant genome; a similar analysis in tomato, but based on different methods to identify regions under selection, suggested about 1% of the genome was under strong selection (Sahu and Chattopadhyay, 2017). That a small portion of the genome is shown to be under selection fits with the observation from QTL studies that in general a small number of QTL control domestication traits (Paterson, 2002). For example, in maize, a small number of large effect genetic variants appear

responsible for the transition from wild teosinte, while variants with smaller effects were responsible for improvement traits that contribute to standing variation in maize crops (Xue et al., 2016).

Eggplant and other members of the Solanaceae exhibit extensive chromosome level synteny (Wu et al., 2009), so regions under selection can be compared. Indeed, it has already been reported that the QTL on chromosome 2 and 4 discussed above overlap with QTL for similar traits in tomato and pepper (Doganlar et al. 2002). Comparing our regions of selection to QTL maps for eggplant (Doganlar et al. 2002; Frary et al. 2003), it appears that regions on chromosomes 2 and 4 may correspond to regions of the eggplant genome controlling fruit size and on chromosome 6 these regions may correspond to regions of the genome controlling prickliness and/or anthocyanin content. We attempted for test for overlap by looking for molecular markers closest to the QTL peaks in the Doganlar et al. (2002) study, and we did not find these markers in our selective sweeps (with one exception; see results), however sometimes the sweep and QTL region appeared close, and given that the QTL sometimes spanned 5-15 cM, there is possible overlap, even if the peaks do not coincide. We also note that the landraces are morphologically diverse and so we might not expect our comparison (wild versus domesticated) to identify e.g. loci associated with fruit colour or shape, given these are very variable in the landraces.

These regions collectively contained 1430 annotated genes, of which ca. 26% had a hit in the Arabidopsis genome and 78% had a hit in the tomato genome. GO enrichment analysis can reveal patterns in selection pressures and can be particularly useful when processing large lists of candidate genes. Using this, we found genes with orthologues related to photosynthesis were significantly over-represented in the regions showing a signal of selection. Photosynthesis is a pathway that has been identified as under selection in a number of crops including tomato (Koenig et al., 2013), other fruit crops (Cao et al., 2014, Li et al., 2019), and non-fruit crops (Pujol et al., 2008, Akakpo et al., 2017), and that has been suggested as a target for crop improvement (Long et al., 2015).

There is no consistent pattern in the change in photosynthetic rate in domesticated species relative to their wild relatives. Increase in photosynthetic rate was observed in domesticated rice (Cook and Evans, 1983), soybean (Li et al., 2013), and cassava (Pujol et al., 2008), whereas photosynthetic rate and photosynthetic pigments were found to be reduced in domesticated yam relative to the wild progenitor (Padhan and Panda, 2018). As domestication often involves a move from a nutrient poor to a nutrient rich habitat, an increase in photosynthetic rate can allow a domesticated species to grow more rapidly in a more protected and nutrient rich environment provided by human cultivation (Pujol et al., 2008). Photosynthetic capacity has been identified as one of the most important targets for increasing crop yield (Long et al., 2015) but has largely been under represented in crop improvement, perhaps due to the complexity of the genetics and the molecular mechanisms controlling traits related to photosynthesis (Mathan et al., 2016). As far as we are aware, photosynthetic rate has not been measured in wild and domesticated eggplant. Quantifying this will be an important next step to understanding the evolution of the photosynthetic pathway during domestication in eggplant.

It could also be that selection on genes related to photosynthesis results from selection on fruit ripening, a process in all fruit crops which involves several pathways that create appealing and edible

fruit. In cultivated tomato, the presence of chloroplasts has been shown to be responsible for the green colour in under ripe fruit (Cocaliadis et al., 2014). A distinguishing feature between wild and domesticated eggplants is that the fruit of wild eggplants retains its green colour throughout ripening, until turning yellow when over ripe, while cultivated eggplants typically ripen to a characteristic purple (but also green, yellow and white fruits are found, especially in landraces). Genes involved in chloroplast biosynthesis under selection during domestication may be responsible for fruit colour change during ripening in eggplants as well as tomato.

Several other over-represented GO terms we identified related to traits which may have diverged during domestication, including a suite of terms related to response to the environment, including osmotic and salt stress, as well as terms potentially related to growth and vigour, including response to hormones and regulation of gene expression and metabolism. Further, the subset of the genes without identifiable orthologues may be of importance to the domestication of eggplant, but we cannot ascribe any sort of function to them at this time. These could be targeted in the future using further *in silico* (Sadowski and Jones, 2009, Galperin and Koonin, 2014) or *in vivo* techniques (Alberts et al., 2015).

We were also limited in the scale of our analysis because a number of 'wild' eggplants were previously identified as feral admixed individuals (Page et al., 2019a). In the future it will be important to expand our survey to include more true wild eggplants allowing a more robust analysis of the genomics of eggplant domestication.

CONCLUSIONS

Sequencing the transcriptome is an effective way of reducing the complexity of the genome, making genome-scale analysis more computationally and cost efficient. However, this approach also comes with the caveat that only expressed genes will be sequenced; for example, in our recent work we identified targets of selection on a gene-by-gene basis using transcriptome sequencing (Page et al. 2019a), which would preclude the sequence analysis of non-expressed or differentially expressed loci. In the present study, by mapping the expressed genes to the genome we overcome this shortfall by identifying SNPs exhibiting signatures of selection (from transcriptome sequencing) and then interrogate the entire genomic region for candidate genes. Some of these signals of selection appear to overlap with previously identified domestication-related QTL and we show that genes in these regions are enriched for functions related to photosynthesis, which is a relatively under-recognised and under-studied domestication trait that has been linked to yield increase in cassava and fruit ripening in tomato.

DATA AVAILABILITY

RNA-seq data are available from the NCBI SRA under accession numbers SRR8736626-SRR8736638, SRR8736644-SRR8736648, SRR8736652, and SRR8736653

FUNDING

This work was supported by a studentship from the University of Southampton for A.M.L.P.

ACKNOWLEDGEMENTS

We would like to thank the staff of the Iridis HPC Facility at the University of Southampton for their computational support throughout. We are especially grateful to Marie-Christine Daunay who has provided significant insight into eggplant domestication in general to us.

REFERENCES

AKAKPO, R., SCARCELLI, N., CHAIR, H., DANSI, A., DJEDATIN, G., THUILLET, A. C., RHONE, B., FRANCOIS, O., ALIX, K. & VIGOUROUX, Y. 2017. Molecular basis of African yam domestication: analyses of selection point to root development, starch biosynthesis, and photosynthesis related genes. *Bmc Genomics,* 18.

ALBERTS, B., JOHNSON, A., LEWIS, J., MORGAN, D., RAFF, M., ROBERTS, K. & WALTER, P. 2015. *Molecular Biology of the Cell,* New York, USA, Garland Science.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M., SHERLOCK, G. & GENE ONTOLOGY, C. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics,* 25**,** 25-29.

BARCHI, L., LANTERI, S., PORTIS, E., VALE, G., VOLANTE, A., PULCINI, L., CIRIACI, T., ACCIARRI, N., BARBIERATO, V., TOPPINO, L. & ROTINO, G. L. 2012. A RAD tag derived marker based eggplant linkage map and the location of QTLs determining anthocyanin pigmentation. *Plos One,* 7.

BARCHI, L., PIETRELLA, M., VENTURINI, L., MINIO, A., TOPPINO, L., ACQUADRO, A., ANDOLFO, G., APREA, G., AVANZATO, C., BASSOLINO, L., COMINO, C., DAL MOLIN, A., FERRARINI, A., MAOR, L. C., PORTIS, E., REYES-CHIN-WO, S., RINALDI, R., SALA, T., SCAGLIONE, D., SONAWANE, P., TONONI, P., ALMEKIAS-SIEGL, E., ZAGO, E., ERCOLANO, M. R., AHARONI, A., DELLEDONNE, M., GIULIANO, G., LANTERI, S. & ROTINO, G. L. 2019. A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. *Scientific Reports,* 9.

BARCHI, L., TOPPINO, L., VALENTINO, D., BASSOLINO, L., PORTIS, E., LANTERI, S. & ROTINO, G. L. 2018. QTL analysis reveals new eggplant loci involved in resistance to fungal wilts. *Euphytica,* 214.

BOLGER, A., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

BROADINSTITUTE. 2019. *Picard Tools* [Online]. Available: http://broadinstitute.github.io/picard [Accessed Jan 2019].

CAO, K., ZHENG, Z. J., WANG, L. R., LIU, X., ZHU, G. R., FANG, W. C., CHENG, S. F., ZENG, P., CHEN, C. W., WANG, X. W., XIE, M., ZHONG, X., WANG, X. L., ZHAO, P., BIAN, C., ZHU, Y. L., ZHANG, J. H., MA, G. S., CHEN, C. X., LI, Y. J., HAO, F. G., LI, Y., HUANG, G. D., LI, Y. X., LI, H. Y., GUO, J., XU, X. & WANG, J. 2014. Comparative population genomics reveals the domestication history of the peach, *Prunus persica*, and human influences on perennial fruit crops. *Genome Biology,* 15.

CATCHEN, J., HOHENLOHE, P. A., BASSHAM, S., AMORES, A. & CRESKO, W. A. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology,* 22**,** 3124-3140.

CHANG, C. C., CHOW, C. C., TELLIER, L., VATTIKUTI, S., PURCELL, S. M. & LEE, J. J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience,* 4.

CHAPMAN, M. A. (ed.) 2019. *The Eggplant Genome,* Springer Nature Switzerland AG: Springer International Publishing.

COCALIADIS, M. F., FERNANDEZ-MUNOZ, R., PONS, C., ORZAEZ, D. & GRANELL, A. 2014. Increasing tomato fruit quality by enhancing fruit chloroplast function. A double-edged sword? *Journal of Experimental Botany,* 65**,** 4589-4598.

COOK, M. G. & EVANS, L. T. 1983. Some physiological aspects of the domestication and improvement of rice (*Oryza* spp). *Field Crops Research,* 6**,** 219-238.

DAVIDAR, P., SNOW, A. A., RAJKUMAR, M., PASQUET, R., DAUNAY, M. C. & MUTEGI, E. 2015. The potential for crop to wild hybridization in eggplant (*Solanum melongena*; Solanaceae) in southern India. *American Journal of Botany,* 102**,** 129-139.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2012. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

DOGANLAR, S., FRARY, A., DAUNAY, M. C., LESTER, R. N. & TANKSLEY, S. D. 2002. Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant. *Genetics,* 161**,** 1713-1726.

DU, Z., ZHOU, X., LING, Y., ZHANG, Z. H. & SU, Z. 2010. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research,* 38**,** W64-W70.

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.

GALPERIN, M. Y. & KOONIN, E. V. 2014. Comparative genomics approaches to identifying functionally related genes. *Algorithms for Computational Biology,* 8542**,** 1-24.

HALL, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series,* 41**,** 95-98.

HOLM, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics,* 6**,** 65-70.

JANSSON, S. 1999. A guide to the *Lhc* genes and their relatives in *Arabidopsis*. *Trends in Plant Sciences*, 4, 236–240.

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.

KOENIG, D., JIMENEZ-GOMEZ, J. M., KIMURA, S., FULOP, D., CHITWOOD, D. H., HEADLAND, L. R., KUMAR, R., COVINGTON, M. F., DEVISETTY, U. K., TAT, A. V., TOHGE, T., BOLGER, A., SCHNEEBERGER, K., OSSOWSKI, S., LANZ, C., XIONG, G., TAYLOR-TEEPLES, M., BRADY, S. M., PAULY, M., WEIGEL, D., USADEL, B., FERNIE, A. R., PENG, J., SINHA, N. R. & MALOOF, J. N. 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.

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.

LI, X. L., LIU, L., MING, M. L., HU, H. J., ZHANG, M. Y., FAN, J., SONG, B. B., ZHANG, S. L. & WU, J. 2019. Comparative transcriptomic analysis provides insight into the domestication and improvement of pear (*P. pyrifolia*) fruit. *Plant Physiology,* 180**,** 435-452.

LI, Y.-H., ZHAO, S.-C., MA, J.-X., LI, D., YAN, L., LI, J., QI, X.-T., GUO, X.-S., ZHANG, L., HE, W.-M., CHANG, R.-Z., LIANG, Q.-S., GUO, Y., YE, C., WANG, X.-B., TAO, Y., GUAN, R.-X., WANG, J.-Y., LIU, Y.-L., JIN, L.-G., ZHANG, X.-Q., LIU, Z.-X., ZHANG, L.-J., CHEN, J., WANG, K.-J., NIELSEN, R., LI, R.-Q., CHEN, P.-Y., LI, W.-B., REIF, J. C., PURUGGANAN, M., WANG, J., ZHANG, M.-C., WANG, J. & QIU, L.-J. 2013. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *Bmc Genomics,* 14.

LONG, S. P., MARSHALL-COLON, A. & ZHU, X. G. 2015. Meeting the global food demand of the future by engineering crop photosynthesis and yield potential. *Cell,* 161**,** 56-66.

MATHAN, J., BHATTACHARYA, J. & RANJAN, A. 2016. Enhancing crop yield by optimizing plant developmental features. *Development,* 143**,** 3283-3294.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research,* 20**,** 1297-1303.

MIYATAKE, K., SAITO, T., NUNOME, T., YAMAGUCHI, H., NEGORO, S., OHYAMA, A., WU, J., KATAYOSE, Y. & FUKUOKA, H. 2020. Fine mapping of a major locus representing the lack of prickles in eggplant revealed the availability of a 0.5-kb insertion/deletion for marker-assisted selection. *Breeding science,* 70**,** 438-448.

PADHAN, B. & PANDA, D. 2018. Variation of photosynthetic characteristics and yield in wild and cultivated species of yams (*Dioscorea* spp.) from Koraput, India. *Photosynthetica,* 56**,** 1010-1018.

PAGE, A., GIBSON, J., MEYER, R. S. & CHAPMAN, M. A. 2019a. Eggplant domestication: pervasive gene flow, feralization, and transcriptomic divergence. *Molecular Biology and Evolution,* 36**,** 1359-1372.

PAGE, A. M. L., DAUNAY, M. C., AUBRIOT, X. & CHAPMAN, M. A. 2019b. Domestication of eggplants: a phenotypic and genomic insight. *In:* CHAPMAN, M. A. (ed.) *Eggplant Genome.* Cham: Springer International Publishing Ag.

PATERSON, A. H. 2002. What has QTL mapping taught us about plant domestication? *New Phytologist,* 154**,** 591-608.

POPELKOVA, H. & YOCUMA, C. F. 2011. PsbO, the manganese-stabilizing protein: Analysis of the structure–function relations that provide insights into its role in photosystem II. *Journal of Photochemistry and Photobiology B: Biology*, 104, 179-190.

PORTIS, E., BARCHI, L., TOPPINO, L., LANTERI, S., ACCIARRI, N., FELICIONI, N., FUSARI, F., BARBIERATO, V., CERICOLA, F., VALÈ, G. & ROTINO, G. L. 2014. QTL Mapping in Eggplant Reveals Clusters of Yield-Related Loci and Orthology with the Tomato Genome. *PLoS ONE,* 9**,** e89499.

PUJOL, B., SALAGER, J. L., BELTRAN, M., BOUSQUET, S. & MCKEY, D. 2008. Photosynthesis and leaf structure in domesticated cassava (Euphorbiaceae) and a close wild relative: Have leaf photosynthetic parameters evolved under domestication? *Biotropica,* 40**,** 305-312.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. W., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics,* 81**,** 559-575.

ROSS-IBARRA, J., MORRELL, P. L. & GAUT, B. S. 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.

SADOWSKI, M. I. & JONES, D. T. 2009. The sequence-structure relationship and protein function prediction. *Current Opinion in Structural Biology,* 19**,** 357-362.

SAHU, K. K. & CHATTOPADHYAY, D. 2017. Genome-wide sequence variations between wild and cultivated tomato species revisited by whole genome sequence mapping. *Bmc Genomics,* 18.

TEAM, R. C. 2015. *R: a language and environment for statistical computing* [Online]. Vienna, Austria: R Foundation for Statistical Computing. Available: http://www.R-project.org/ [Accessed].

TOPPINO, L., BARCHI, L., LO SCALZO, R., PALAZZOLO, E., FRANCESE, G., FIBIANI, M., D'ALESSANDRO, A., PAPA, V., LAUDICINA, V. A., SABATINO, L., PULCINI, L., SALA, T., ACCIARRI, N., PORTIS, E., LANTERI, S., MENNELLA, G. & ROTINO, G. L. 2016. Mapping quantitative trait loci affecting biochemical and morphological fruit properties in eggplant (*Solanum melongena* L.). *Frontiers in Plant Science,* 7**,** 256.

WEIR, B. S. 1996. *Genetic data analysis II: methods for discrete population genetic data,* Sunderland, Massachusetts, Sinauer Associates, Inc.

WU, F. N., EANNETTA, N. T., XU, Y. M., DURRETT, R., MAZOUREK, M., JAHN, M. M. & TANKSLEY, S. D. 2009. A COSII genetic map of the pepper genome provides a detailed picture of synteny with tomato and new insights into recent chromosome evolution in the genus *Capsicum*. *THEORETICAL AND APPLIED GENETICS,* 118**,** 14.

XUE, S., BRADBURY, P. J., CASSTEVENS, T. & HOLLAND, J. B. 2016. Genetic architecture of domestication-related traits in maize. *Genetics,* 204**,** 99-+.

TABLE AND FIGURE LEGENDS:

Table 1 – Samples from which RNA-seq data was used. [1] Accessions names beginning 'MM' are from INRA, 'Meyer' indicates collection of R. Meyer (UCLA), 'PI' are from the USDA, 'TS' and 'S' are from the AVRDC, 'ARUM' indicates a landrace from Amishland seeds. [2] phenotypic information is given where available.

Table 2 – Genomic regions of low diversity (low π in domesticated) or high differentiation (high F_{ST}) in the comparison of the wild and domesticated eggplant populations.

Figure 1 – Cumulative percentage graph of haplotype block size. The point at which the rate of increase of cumulative percentage declines is marked on the x and y axis.

Figure 2 – Chromosomal distributions of SNPs and outliers represented through Manhattan plots of -$\log_{10}$(P) for F_{ST} and π. Significant loci after Holm-Bonferroni correction are beyond the dotted line. Centre ideogram chromosomes show gene density. Plotted using KaryoploteR in R (Gel and Serra, 2017).

Table 1

Accession[1]	Population	SRA accession	N reads	Country of Origin	Spines[2]	Fruit colour[2]	Fruit shape[2]
MM12137	domesticated	SRR8736626	19623124	India			
MM1712	domesticated	SRR8736644	11685753	China	no	purple	elongated
MM1791	domesticated	SRR8736653	15531766	Vietnam	yes (calyx)	purple	round
Mey319	domesticated	SRR8736652	26896589	China	no	purple	round
PI241594	domesticated	SRR8736633	18493989	Taiwan			
Arum	domesticated	SRR8736631	21310738	India			
MM0609	domesticated	SRR8736630	12326452	India	no	purple	round
MM0673	domesticated	SRR8736638	22010086	India			
MM10439	domesticated	SRR8736627	23224957	Maldives			
S00255B	domesticated	SRR8736636	22806997	India			
S00392	domesticated	SRR8736637	22688619	India	no	purple	elongated
MM12391	domesticated	SRR8736632	23184716	Malaysia	no	purple	round
MM12454	domesticated	SRR8736635	23079041	Indonesia			
MM1290	domesticated	SRR8736634	14460317	Philippines	no	purple	elongated
MM1547	domesticated	SRR8736629	15435197	Malaysia	no		
PI470273	domesticated	SRR8736628	22634687	Indonesia	no	purple	elongated
MM0669	wild	SRR8736648	22235038	India	yes	green	round
MM0675	wild	SRR8736647	18680378	India	yes	green, variegated	round
PI381155	wild	SRR8736645	18387079	India			
MM0686	wild	SRR8736646	24282614	Indonesia			

Table 2

Region name	Test	N sig SNPs	Chr	Position	N genes
1_1	π	3	1	107964200 - 115824939	73
2_1	π	2	2	13892778 - 22201207	54
2_2	π	6	2	60392355 - 68261479	39
2_3	F_{ST}	6	2	69739198 - 77677099	285
4_1	π	5	4	42196297 - 50056466	53
6_1	π	5	6	79014216 - 86874666	179
6_2	F_{ST}	2	6	92198368 - 100135870	526
9_1	π and F_{ST}	16	9	5546492 - 13498317	221

Figure 1

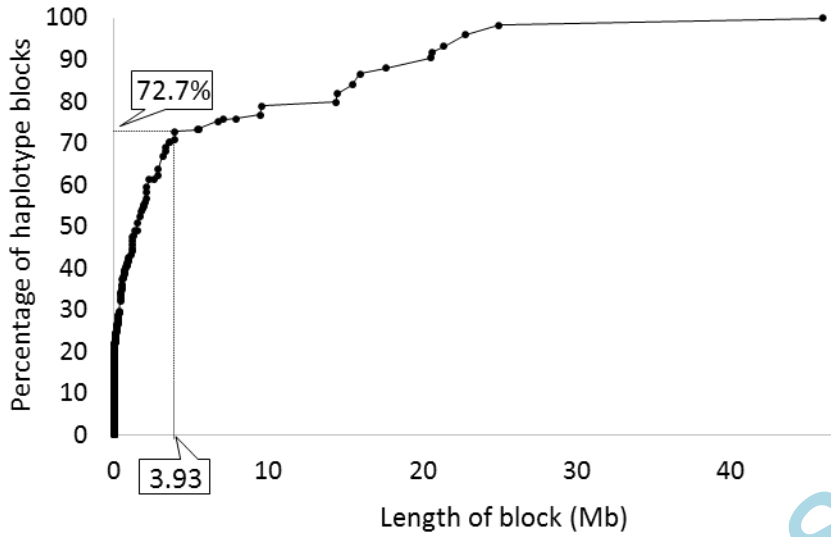

Figure 2