# Comparative analyses of the molecular footprint of domestication in three Solanaceae species : eggplant, pepper and tomato

Stéphanie Arnoux

▶ **To cite this version:**

Stéphanie Arnoux. Comparative analyses of the molecular footprint of domestication in three Solanaceae species : eggplant, pepper and tomato. Agricultural sciences. Université d'Avignon, 2019. English. NNT : 2019AVIG0351 . tel-02185095v2

**HAL Id: tel-02185095**

**https://theses.hal.science/tel-02185095v2**

Submitted on 17 Jul 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Avignon Université
*École Doctorale 536 Agrosciences et Sciences*

INRA Centre de recherche PACA
*UR 1052, Génétique et Amélioration des Fruits et Légumes*

# THÈSE

# Étude comparée des traces génétiques de la domestication chez trois Solanacées : l'aubergine, le piment et la tomate

Présentée à Avignon Université
pour obtenir le grade de

Docteur en Sciences

*Spécialité : Biologie*

Soutenue le 21 février 2019
par

Stéphanie Arnoux

*Encadrée par :*

Dr. Mathilde Causse (DR, HDR, INRA Avignon) — *Directeur de thèse*
Dr. Christopher Sauvage (CR, INRA Avignon) — *Co-encadrant*

*Devant le jury composé de :*

Dr. Maud Tenaillon (DR, HDR, CNRS Gif/Yvette) — Rapporteur
Dr. Joëlle Ronfort (DR, HDR, INRA Montpellier) — Rapporteur
Dr. Concetta Burgarella (Chercheur, CIRAD Montpellier) — Examinateur
Dr. Jérémy Clotault (MC, Université d'Angers) — Examinateur

# Résumé

La domestication des plantes a débuté il y a quelques milliers d'années quand les hommes se sont sédentarisés. Ils ont sélectionné les plantes sauvages portant des caractères phénotypiques d'intérêt pour la consommation et production humaine. Ce processus évolutif a par conséquent modifié le patrimoine génétique des espèces domestiquées. Cette thèse se penche sur les traces génétiques induites par la domestication chez trois espèces de Solanacées : l'aubergine (*Solanum melongena*), le piment (*Capsicum annuum*) et la tomate (*S. lycopersicum*). En effet, si les caractères phénotypiques des plantes cultivées ont été sélectionnés depuis des milliers d'années, les conséquences moléculaires d'une telle sélection restent peu étudiées à l'échelle du génome. Cette étude est basée sur des données de diversité et d'expression de gènes (RNAseq). En utilisant des méthodes comparatives entre des variétés cultivées et leurs espèces sauvages apparentées, j'ai étudié, à l'échelle intra-spécifique, d'une part les histoires démographiques de chacune des espèces, et d'autre part les changements de diversité nucléotidique et d'expression des gènes dus à la domestication. La comparaison de ces trois événements indépendants de domestication, offre l'opportunité de décrypter les changements génétiques qui convergent chez ces trois espèces lors du processus de sélection humaine.

Suite à une introduction qui pose le cadre de cette étude et présente l'état de l'art, le premier chapitre, s'inscrit dans un ouvrage portant sur la génomique des populations d'espèces modèles. Il propose une synthèse des connaissances accumulées en plus d'un siècle de recherche sur l'espèce modèle qu'est la tomate (*S. lycopersicum*). Ce chapitre permet également de compléter le contexte scientifique dans lequel cette thèse s'inscrit, notamment, en retraçant l'importance que les espèces sauvages apparentées ont eu dans l'amélioration de l'adaptabilité des variétés cultivées actuelles.

L'hypothèse du deuxième chapitre révèle la convergence des changements démographiques entre les trois espèces malgré leurs événements indépendants de domestication. L'étude comparée d'inférences de scénarios démographiques a permis de reconstruire l'histoire démographique de chaque espèce cultivée. Ces inférences ont aussi facilité l'estimation des paramètres tels que les flux migratoires entre les espèces sauvages et cultivées, la force des goulots d'étranglement liés à l'intensité de la sélection humaine et la durée des événements de domestication. Ce chapitre permet de démontrer que les changements démographiques liés à la domestication dépendent de l'état de sympatrie ou d'allopatrie des variétés cultivées avec leurs sauvages apparentées. Les connaissances quant à la datation des événements de domestication de nos trois espèces restent très faibles, et les inférences ont permis d'établir des estimations de durée de domestication relativement précise. Ces nouvelles connaissances apportent une plus-value à cette étude pour nos trois espèces et nous invitent à s'interroger sur les différents compartiments du génome qui ont été sélectionnées et modifiées lors de la domestication.

Le troisième chapitre teste l'hypothèse d'une convergence évolutive des changements moléculaires, notamment transcriptionnels, induits par la domestication et l'amélioration moderne. La comparaison des variétés cultivées à leurs espèces sauvages apparentées permet d'évaluer la convergence des mécanismes de régulation et d'adaptation liés à la domestication. C'est en testant la corrélation entre les traces génétiques (diversité nucléotidique) de sélection et les changements

d’expression des gènes observés chez les variétés cultivées que l’hypothèse de départ a été validée. Cette analyse montre que la domestication, au-delà même de changements nucléotidiques, a modifié l’expression des gènes chez les trois espèces. L’analyse des gènes orthologues des espèces a confirmé que la domestication a sélectionné des gènes liés aux phénotypes de développement des fruits et la croissance de la plante alors qu’elle avait, au contraire, contre-sélectionné des gènes liés à la défense des plantes et à leur capacité à tolérer des stress environnementaux.

Enfin, en discussion, je réalise un bilan sur mon projet qui apporte de nombreuses preuves de convergence dues à la domestication et des connaissances utiles pour l’étude de l’histoire des Solanacées. De surcroit, des perspectives d’analyses complémentaires sur la liste de nombreux gènes candidats affectés par la domestication, offrent un potentiel de transversalité, pour l’amélioration des variétés cultivées et pour l’étude plus approfondie des conséquences biologiques et évolutives de la domestication.

## Summary

Domestication started thousand years ago when human shifted from hunter-gatherer to agrarian societies. They started selecting wild plants for phenotypes related to consumption and yield. This evolutionary process induced changes in the gene pool of domesticated plants. This thesis focuses on genetic footprints induced by domestication within a trio of *Solanaceae* species: the eggplant (*Solanum melongena*), the pepper (*Capsicum annuum*) and the tomato (*S. lycopersicum*). Crop plants have been selected for thousand years on phenotypic traits, but the molecular consequences of such selection remain unknown at the genome-wide scale. The study was performed on a RNAseq data set; using comparative methods between crops and their wild relatives, I studied, at the intra-specific scale, the demographic history, and, both the nucleotide diversity and the gene expression changes due to domestication. Comparing these three independent events of domestication, is a great opportunity to decipher the interspecific genetic changes, converging for the three species, during the human selection process.

The first chapter is a book chapter about population genomics in model species. It details the state of art of hundred years of research on tomato as model species (*S. lycopersicum*). Tomato is a model species in genetics, as well as in population genomics thanks to the important collection of genomic data that have been accumulating over years. Tomato has the strongest economic importance within the trio of studied species. By highlighting the importance of crop wild relative species for adaptability improvement of modern cultivars, this chapter describes the scientific context of this thesis work.

The two next chapters are following these researches and show the importance to both conserve and study the crop wild relative species.

In the second chapter, I hypothesize that demographic changes within the three species experience a convergence, despite their independent domestication events. The comparative study of demographic inferences allows the reconstruction of each domesticated species demographic history. Theses inferences facilitate the parameter estimations such as the migration rate between crop and wild, the bottleneck strength paired with the human selection and the duration of the domestication events. This chapter reveals a common bottleneck phenomenon as well as migration rate dependent to the allopatric or sympatric state of the crops with their wild relatives. Knowledge concerning the domestication events dating, for each of the three species, remain poorly studied and this thesis work discloses relative domestication time durations.

These new insights bring valuable knowledge to the three species and induce a questioning on the different genome parts that are selected and modified through domestication.

The third chapter, test the hypothesis of a convergent evolution of molecular changes, especially transcriptional, induced by domestication and modern breeding. The comparative analysis of crop plants and their wild relatives assesses the convergence of regulation and adaptation mechanisms due to domestication. By testing the correlation between the selection footprints on genes and the gene expression changes in crop compared to their wild relative species, the previous hypothesis was confirmed. This analysis implies that domestication modified gene expression in the three species beyond only nucleotide polymorphisms. The ortholog analysis of our species genes,

confirmed that domestication facilitated the fruit development and plant growth but relaxed selective pressure on genes of plant defense and environmental stresses tolerance.

Demonstrating demographic changes and molecular footprints of domestication, my PhD thesis highlights several proofs of convergence. It offers estimations of duration of domestication that are valuable for the study of agrarian history of *Solanaceae*. It supplies numerous candidate genes impacted by domestication, with transversal potential (orthologs in the three species), that could improve greatly the modern cultivars. Such genes could be thoroughly analyzed to improve the common understanding of biological and evolution consequences of domestication in *Solanaceae*.

# Table of Contents


***Table of Contents*** ..... ***- 5 -***

***INTRODUCTION*** ..... ***- 9 -***

***I. Scientific Context*** ..... ***- 11 -***

**a. Domestication** ..... **- 11 -**

i. Domestication definition ..... - 11 -

ii. The four stages of domestication ..... - 12 -

**b. Phenotypic and genetic consequences of domestication** ..... **- 14 -**

**c. From gene to genome: -omics footprints of domestication** ..... **- 17 -**

i. Domestication induced a genomic diversity reduction ..... - 17 -

ii. Changes at the transcriptome level induced by domestication ..... - 18 -

iii. Metabolome changes at the genome-wide scale ..... - 19 -

**d. Crop wild relatives, the source of potential for improvement** ..... **- 20 -**

**e. Potential of demographic inferences to decipher domestication** ..... **- 22 -**

***II. Focus on the study systems: Solanaceae*** ..... ***- 26 -***

**a. The Solanaceae family** ..... **- 26 -**

i. Economic importance ..... - 27 -

ii. Scientific recognition ..... - 28 -

**a. Eggplant history** ..... **- 30 -**

i. Taxonomy and species history ..... - 30 -

ii. Genetic resources ..... - 32 -

iii. Molecular markers and genome mapping ..... - 33 -

**b. Pepper history** ..... **- 34 -**

i. Taxonomy and species history ..... - 34 -

ii. Genetic resources ..... - 36 -

iii. Molecular markers and genome mapping ..... - 37 -

**c. Tomato history** ..... **- 39 -**

***III. Scientific questions and hypothesis of the thesis*** ..... ***- 40 -***

***MATERIALS AND METHODS*** ..... ***- 43 -***

**a. Data available before the start of the PhD project** ..... **- 43 -**

**b. Choice of plant accessions** ..... **- 44 -**

**c. Preparation of the biological material** ..... **- 47 -**

**d. Alignment of the RNAseq data set** ..... **- 47 -**

**e. Demographic inference modeling** ..... **- 48 -**

**f. Gene expression analyses** ..... **- 49 -**

**g. Complementary details on the bioinformatic workflow (p 51-54)** ..... **- 49 -**

i. Common bioinformatic workflow to both chapter analyses ..... - 49 -

ii. Comparing inference modeling ..... - 49 -

iii. Transcriptomic, ortholog, gene ontology and nucleotide diversity analyses ..... - 50 -

*CHAPTER 1 ..... - 57 -*

*Progress and prospects of population genomics in major crop plants - Tomato population genomics ..... - 57 -*

**Abstract - 61 -**

**Introduction - 63 -**

**Part I: How tomato became the model plant for vegetables - 64 -**
1. Tomato history, from past to modern era ..... - 64 -
2. Towards the reference genome of tomato and databases ..... - 65 -
3. Genome and transcriptome sequencing of crop wild relative species ..... - 66 -

**Part II: Tomato as a model for Molecular Evolution - 67 -**
1. Original organization of the Tomato clade ..... - 67 -
2. Modern phylogeny and taxonomy of the Tomato clade ..... - 68 -
3. Ecological Genomics of the tomato crop and its wild relatives ..... - 70 -
4. Genomic footprints of domestication and modern breeding stages ..... - 76 -

**Part III: Population genomics to sustain modern breeding - 79 -**
1. Introgressions from crop wild relative species improved the crop tomato ..... - 80 -
2. Dissecting the genetic architecture of agronomical traits ..... - 81 -
3. Molecular bases of trait diversification ..... - 82 -
4. Breeding shaped the genetic structure of modern cultivars ..... - 82 -
5. Genome-wide association approach extended the knowledge of the genetic architecture of agronomical traits - 84 -

**Part IV: Prospects for future research - 87 -**
a. Towards a pan-genome in tomato ..... - 87 -
b. Modelling of demographic history and ecological niche ..... - 87 -
c. Adapting the tomato crop to climate change using genomic approaches ..... - 89 -
d. Implementing genome-wide based Genomic Selection ..... - 89 -

Further Major Readings We Recommend: ..... - 92 -

*CHAPTER 2 ..... - 93 -*

*Demographic inferences reveal a convergence of domestication in Solanaceae ..... - 93 -*

**Abstract - 97 -**

**1. Introduction - 99 -**

**2. Materials and Methods - 100 -**
Plant Materials and RNA sequencing ..... - 100 -
Quality control, reads alignment and variant calling ..... - 101 -
PCA and unfolded allele frequency spectrum ..... - 102 -
Demographic inferences ..... - 102 -

**3. Results - 103 -**
Biological material and high-throughput sequencing ..... - 103 -
Genetic structure and Allele frequency spectrum ..... - 104 -
Demographic inferences and best scenario choice ..... - 108 -
Parameter estimates and bootstrap of each species' best model ..... - 108 -

**4. Discussion - 110 -**
Domestication footprints on Solanaceae genomes: impact of the artificial selection ..... - 110 -
A convergent bottleneck revealed by the comparative analysis of three Solanaceae species ..... - 111 -
Timing of the different stages in the domestication process ..... - 112 -
Conclusion ..... - 114 -

**Acknowledgements** - 114 -

**Data Accessibility Statement** - 115 -

**Author Contributions** - 115 -

**Supplementary Figures** - 116 -

**Supplementary Tables** - 117 -

***CHAPTER 3*** ..... ***- 119 -***

***Domestication footprints reveal a convergence of both nucleotide diversity and gene expression in cultivated Solanaceae*** ..... ***- 119 -***

**Abstract** - 123 -

**1. Introduction** - 125 -

**2. Material and Methods** - 126 -

Plant Materials ..... - 126 -
Alignment pipeline ..... - 127 -
Nucleotide diversity ..... - 128 -
Differential expression analyses ..... - 128 -
Annotations and orthology analyses across the three species ..... - 129 -
Statistical analyses ..... - 129 -
Data availability ..... - 130 -

**3. Results** - 130 -

Biological material ..... - 130 -
Identification of genetic diversity shifts ..... - 132 -
Identification of differentially expressed genes ..... - 133 -
Annotations and enrichment analyses ..... - 134 -
Statistical analyses ..... - 137 -

**4. Discussion** - 139 -

**Acknowledgements** - 143 -

**Author contributions** - 144 -

**Supplementary Figures** - 145 -

**Supplementary Tables** - 147 -

***GENERAL DISCUSSION*** ..... ***- 149 -***

a. Choice of the biological material: does a subset of accessions represents a species? ..... - 151 -
b. Advantages and limitations of the likelihood *vs* Bayesian methods for inferring demographic models. ..... - 155 -
c. Crop and wild comparative analyses are powerful to decipher domestication footprints, the case of the PhD work and further outlook. ..... - 157 -
d. Studying evolutionary transcriptomics reveals that domestication induced modifications in different mechanisms of gene expression regulation. ..... - 159 -
e. The next improvement to infer domestication might involve the implementing of if the variation in environmental conditions over time. ..... - 162 -
General conclusion and outlook ..... - 163 -

***BIBLIOGRAPHY*** ..... ***- 167 -***

*APPENDIX 1: Detailed description of the 92 accessions available for the studies..... - 189 -*

*APPENDIX 2: Supplementary tables related to the chapter 2 ..... - 192 -*

*APPENDIX 3: Supplementary tables related to the chapter 3 ..... - 205 -*

*APPENDIX 4: Reviewers comments and major questions for Genome, Biology and Evolution - 219 -*

*ACKNOWLEDGEMENTS..... - 223 -*

*Résumé substantiel de la thèse en français..... - 227 -*

# INTRODUCTION

# I. Scientific Context

Agriculture appeared 45 to 65 Million years ago in Amazonian rainforests when attine ants started to cultivate fungi, depending on this crop for food. The domestication process has been documented to have evolved independently at least 5 times in evolutionary history, such as for the cultivation of fungal species by specific ant, termite and beetle species (Mueller et al. 2005b), but it is human that specialized in domestication by cultivating the greater number of plant species. This thesis work proposes to focus on the domestication process that have a strong importance for many scientific fields such as evolutionary biology, crop science and archeology. Studying such process helps: understanding artificial selection; bringing valuable insights for improvement and breeding effort; and deciphering human cultural and societal history (Meyer and Purugganan 2013).

## a. Domestication

### i. Domestication definition

The domestication definition has been long discussed, essentially regarding who benefited more from the relationship: namely the domesticate (species modified and/or created through the process) or the domesticator (species that induces the phenotypic and genetic modifications) (Rindos 1983; O'Connor 1997; Ervynck et al. 2001). Here the wider definition is considered in which the domesticator has acquired the knowledge necessary to manage the reproduction and the care of the domesticate, in order to obtain sustainable phenotypes of interest. In this definition, and within the given growth conditions controlled by the domesticator, the co-evolutionary interaction benefits to both the domesticate and the domesticator, observing an increase of their fitness (Zeder 2015). This definition focuses more on the relationship than on the genetic or plastic outcomes of it.

The evolutionary transition that is domestication, already inspired Darwin who described it as a great opportunity to study genetic variation, evolution and selection (Darwin 1868). Since then, domestication has been of high interest for the scientific community, willing to learn more on the human impact on domesticated animals and plants. Evolutionary biology was initiated earlier on, by Darwin, when he resolved the question of the diversity of forms of life, by giving an ecological explanation on how natural selection works, at a phenotypic scale, in favor to the fittest form within a specific environment (Darwin 1859). Following this path a neo-Darwinian theory of evolution resolved the genetic scale by formulating how changes in gene frequencies that were driving

evolution were due to mechanisms of mutation, selection and drift (Huxley 1963). The evolutionary process of plant domestication is human-triggered, and the current crop species have been long selected. For example, in the Middle East it started as long as at the early Holocene era (~12,000-11,000 years ago). The selection on plant targeted phenotypes interesting for consumption and agricultural purposes. But, already back in 1886, de Candolle observed the discrepancy that exists within cultivated plants. Even if plants were domesticated, he observed that domestication was following several stages and made the hypothesis that selection on early stages of domestication was more important than selection on cultivar varieties (de Candolle 1886). Following his work, Vavilov (1926) introduced the concept of primary and secondary pools, where primary pools are the first accessions brought under cultivation, and secondary ones being the ones derived for the primary after experiencing selection due to the processes of agriculture.

### ii. The four stages of domestication

The domestication process has been defined as following 4 stages (Meyer and Purugganan 2013) depicted in figure 1 from Gaut *et al.* (2018). These four stages of domestication corroborate with the four degrees of domestication namely wild, semi-domesticated, domesticated and modern cultivar (Clement 1999).

(i) The first stage consists in a protracted period in which the species is separated from its wild progenitor, but remains in its wild environment (Zeder 2015). This stage can be assimilated to foraging as niche construction, it is mainly a management of wild resources favoring phenotypes of interest for consumption but remaining a small-scale plant cultivation for hunter-gatherers (Rowley-Conwy and Layton 2011). Humans modified the average phenotype from the range of variation that remained in the wild population. In the protracted transition, the hypothesis is that artificial selection is weak for a long time (Allaby 2010). Selecting only few wild individuals per generation induced a slow reduction of the genetic variability (Doebley 1989b). The plant is considered as semi-domesticated, it retains enough adaptive potential to survive in the wild but its selected phenotypes would disappear over time in its natural environment (Clement 1999).

(ii) The second stage occurs when human voluntary started the cultivation. This stage is different as humans artificially selected phenotypes for their consumption and farm production but most of all, they started to control the breeding and improve traits across generation (Zeder 2006). By shifting the plant environment to cultivation area, humans ensured the control of seed dispersal,

plant growth and breeding and created a new niche for crops (Fuller and Allaby 2009). While choosing the best adapted varieties to grow in the cultivated landscapes (Harlan 1992), only few genotypes may be domesticated. The pressure of selection induced a genetic bottleneck that strongly reduced the genetic variability, especially in annual plants (Miller and Gross 2011), this is paired with a loss of ecological adaptation.

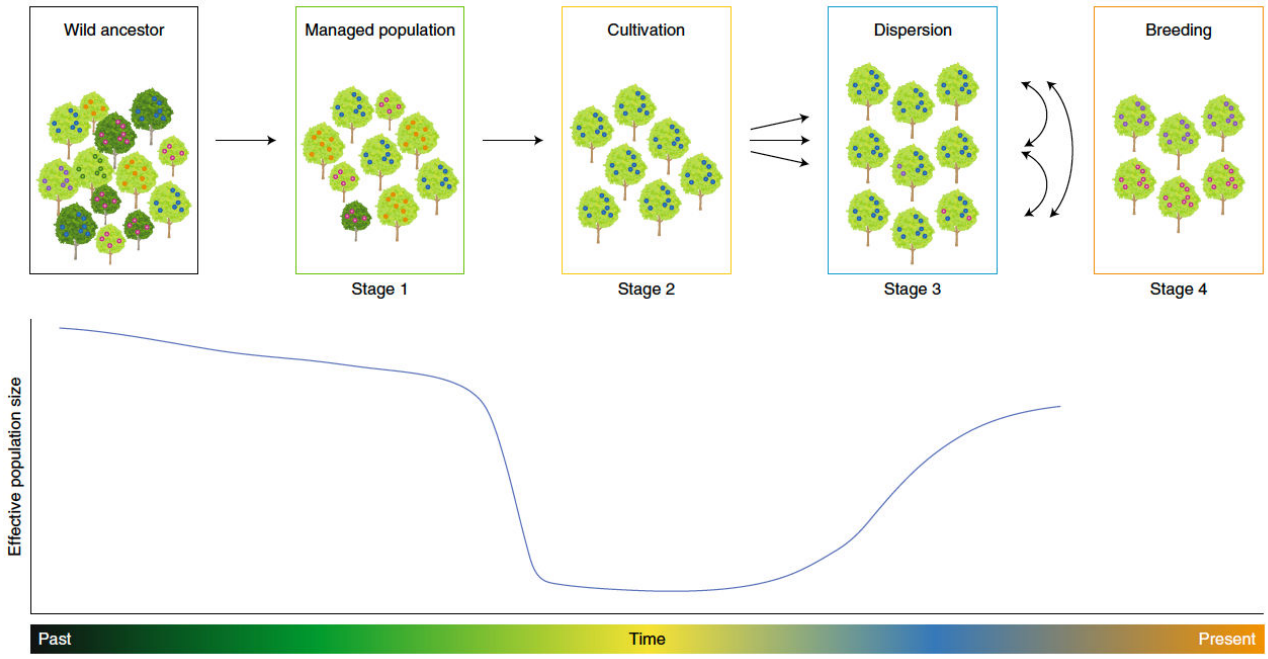



**Figure 1**. Features of demography and selection during plant domestication. A schematic representing four stages of domestication. The far-left population represents wild populations with substantial genetic diversity. The curve below the stages provides an example of population size through the four stages, including a long population decline through stage 1 and an abrupt bottleneck in stage 2, followed by population extension. *Source: Gaut et al. 2018*

(iii) At this third stage of domestication, humans created environments specifically developed to meet the optimal growth requirement of the domesticates in a purpose to enhance the yield and the predictability of production (Zeder 2015). The geographic expansion of domesticates required local adaptations to the new human environments, and it had the effect of increasing the genetic variability and the effective population size. But, even if the domesticated plants experienced an increase in genetic variability, their adaptation to a specific geographical location did not improve

their ability to survive without humans. This domestication stage conferred the plant a status of landraces (Zeder 2006).

(iv) The fourth stage is deliberate breeding, this stage is quite recent, only hundred years ago and is a conscious selection of improved specific crop phenotypes and genotypes. The domesticated plant upgraded to a modern cultivar status, where both phenotypic and genetic variabilities were reduced. When mostly clonally propagated, they were adapted exclusively to intensive monoculture.

### b. Phenotypic and genetic consequences of domestication

**Figure 2**. Several pictures of the wild to crop phenotypic conversion. (A) Teosinte to maize ear: change from a few small, loosely connected seeds with thick fruit cases to a large maize cob with many naked seeds. (B) Loss of shattering in crop rice. (C) Fruit size increase in tomato. (D) Loss of branching in sunflower. *Source: Doebley et al. 2006 & Stetter et al. 2017*

The phenotypic traits selected during domestication that differentiate the crop from their wild relatives are collectively known as 'Domestication syndrome'. These traits are common to all annual crops, from cereals to vegetables, they mostly include gigantism of the harvested part of the plant, reduction of the branching and thorn, and less shattering (Figure 2). Such phenotype-targeted selection induced parallel and convergent selection between crop plants regardless their taxonomic family, and despite the dozen of independent domestication centers differing greatly in respect of geographical location, center size, number and diversity of domesticated species and their respective potential as food sources (Doebley et al. 2006; Larson 2014). To obtain similar phenotypes of interest, a parallel selection involves different Quantitative Traits Loci (QTLs) when a convergent selection impacts the same QTLs or ortholog genes (i.e. homologous genes that derived from the same ancestral gene) across the different crop species (Fuller et al. 2014).

The figure 3 represents examples of protein coding genes that are related to domestication-targeted phenotypes, some of them resulting of parallel selection across plant families such as the shattering that involves *SH1* in Poaceae (Lin et al. 2012) but *Shat 1-5* and *PdH1* in Leguminosae (Sedivy et al. 2017). The seed dormancy is a phenotype with convergent evolutive trajectory implying a unique gene (namely *stay-green G*), for the three plant family: Leguminosae, Solanaceae and Poaceae (Wang et al. 2018). Though it is to mention that most of the literature refers to phenotypic changes induced by domestication as convergent but molecular changes as parallel, whether or not there are on a common gene (Rendón-Anaya and Herrera-Estrella 2018).

In the case of annual plants, such as for *Solanaceae*, the fruit is considerably modified with an important increase in size and shape diversity. For example, in tomato the fruit size increased by a 100x fold in crop, and the underlying QTL *fw2-2* is proven to induce up to 30% of this change (Frary 2000). With extensive research on the domestication of tomato, it appears that modification of fruit shape is controlled by four genes, *SUN* and *OVATE* controlling the elongated shape, and, *LOCULE NUMBER (LC)* and *FASCIATED (FAS)* controlling the locule number and flat shape, respectively (Rodriguez et al. 2011). However, the fruit size and shape are not the only traits that were modified by domestication. Plant growth and architecture were deeply modified during domestication with the fixation of a single amino change in the *SELF-PRUNING* (*SP*) gene that 'determines' the growth of the plant. In the wild, tomato are indeterminate and the crop with *sp* mutation experiences a reduction of leaf number between trusses and a replacement of leaves by flowers and growth stop (Pnueli et al. 1998). This trait is of particular interest for harvesting tomato in open field for industrial purposes (i.e. producing tomato paste and juice). The *SP* gene is ortholog with *CENTRORADIALIS* and

*TERMINAL FLOWER1* genes of *Arabidopsis thaliana* (Bradley et al. 1996, 1997), and both ortholog genes are involved in plant determination. The similar role of these ortholog genes in the two highly differentiated plant families confirms the potential of translational work when focusing on genes targeted by domestication.

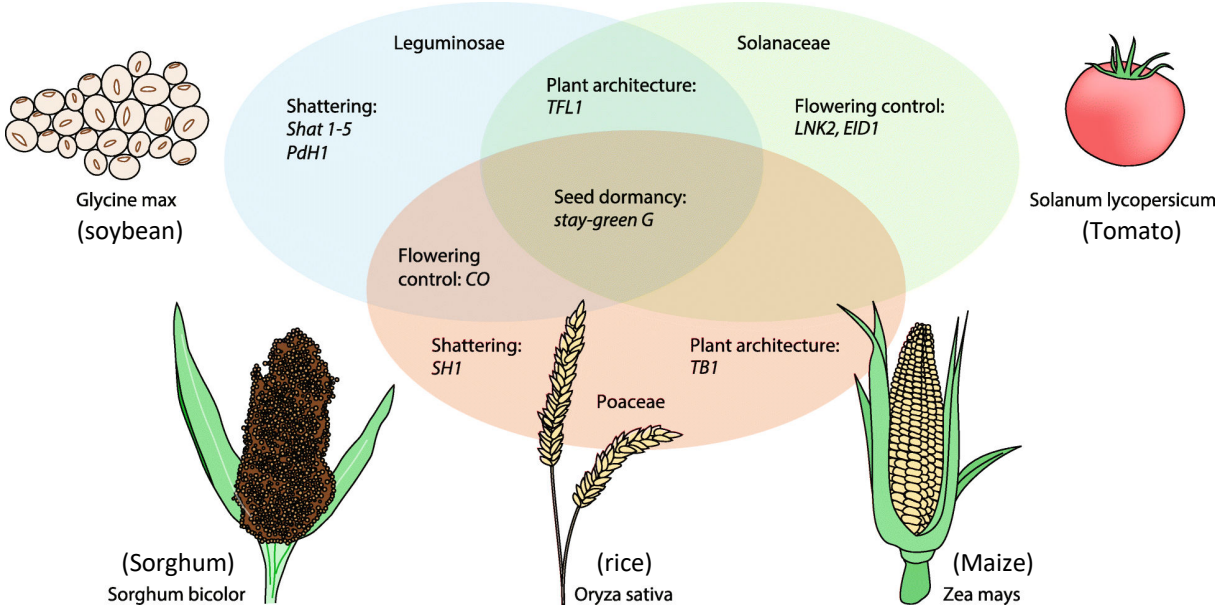



**Figure 3**. Examples of parallel (and convergent) selection of protein-coding genes across plant families. *Source: Rendón-Anaya and Herrera-Estrella 2018*

To find candidate genes related to domestication, two complementary approaches can be used: quantitative genetics and population genetics. Both methods aim to find genetic signature of domestication. Quantitative genetics uses a top-down method to detect the candidate genes associated with a phenotype of interest (especially powerful to detect the major effect genes), whereas population genetics is more of a bottom-up method that focuses on genetic signature of domestication to detect genes that were selected in the crop species and showing an 'outlier behavior' (Ross-Ibarra et al. 2007).

## c. From gene to genome: -omics footprints of domestication

### i. Domestication induced a genomic diversity reduction

With the outbreak of high-throughput genotyping and phenotyping methods, the power of quantitative and population genetic analyses greatly increased. The availability of genome-wide data revealed one of the main unforeseen consequences of domestication: the genome-wide reduction in crop genetic diversity (Doebley et al. 2006). The selection for favorable alleles induced selective sweeps that imprinted the whole genome, as shown in maize (Hufford et al. 2013), rice (Caicedo et al. 2007; Nabholz et al. 2014) and tomato (Koenig et al. 2013; Sauvage et al. 2017). This selection was paired with a relaxation of natural selection on traits that lost importance in the crops (Innan and Kim 2004). And even if directional selection is the main actor in the domestication process, diversifying selection is active on target loci associated with domesticated phenotypes in the common bean (*Phaseolus vulgaris*). Moreover, by favoring selfing as mating system to preserve genetic background across generations, cultivation increased the crop inbreeding. This practice also impacted the recombination rates by reducing the crossover effectiveness in breaking up linkage groups, hence it enhanced the decay of linkage disequilibrium in crops (Ellegren and Galtier 2016). The linkage disequilibrium increased the hitchhiking of neutral genes present in flanking regions of selected genes. The second drawback of such hitchhiking is the accumulation of deleterious mutations as shown in rice cultivars where more than a quarter of the amino acid differences are deleterious (Lu et al. 2006).

As represented in the figure 4*a*, the selection is paired with demographic changes such as a bottleneck that simultaneously induces a reduction of the nucleotide diversity and an increase in accumulation of deleterious mutations, this last phenomenon being named the 'cost of domestication' (Moyers et al. 2018). Even if the genome wide diversity is impacted, the selection is uneven across chromosomes and genes that are selected experience more severe bottleneck than the unselected ones. In this context, scanning a crop population for genetic diversity ($\pi$) or Tajima's D, allows the identification of specific selected regions such as selective sweep (Nielsen 2005; Lai et al. 2010). In tomato, the use of genomic analyses proved that both population and quantitative genetic methods combined were revealing domestication related candidate genes (Lin et al. 2014). A model-based clustering analysis complemented their analysis and provided insights into the different stages of tomato domestication, as represented in the figure 4*b*: stage 2 (*S. lycopersicum*

var. *cerasiforme*), stage 3 (*S. lycopersicum*, big-fruited considered as landraces) and stage 4 (*S. lycopersicum*, selected for processing or fresh market purposes considered as modern cultivars).

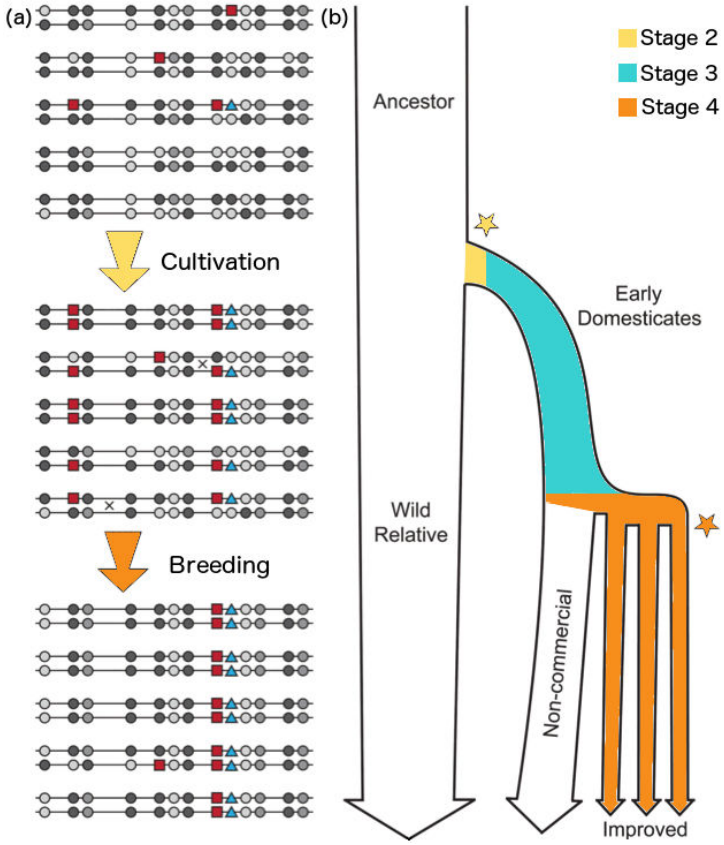



**Figure 4**. Processes of cultivation and breeding.

(*a*) Effects of artificial selection (targeting the blue triangle variant) and linkage disequilibrium on deleterious (red squares) and neutral variants (grey circles, shades represent different alleles).

(*b*) Typical changes in effective population size through domestication. Stars indicate genetic bottlenecks. These dynamics can be reconstructed by examining patterns of genetic diversity in contemporary wild relative, domesticated non-commercial, and improved populations.

*Adapted from: Moyers et al. 2018*

### ii. Changes at the transcriptome level induced by domestication

The changes in nucleotide diversity due to domestication occurred with a rewiring of gene expression levels, as observed in a few species such as maize (Wright 2005; Swanson-Wagner et al. 2012), tomato (Koenig et al. 2013; Sauvage et al. 2017) and common bean (Bellucci et al. 2014). At the transcriptional level, few studies have deeply characterized the parallel changes induced by the domestication across crop species. In common bean, for example, 18% to 26% of the diversity of gene expression was lost through domestication. Not only are the genes differentially expressed, but 74% of them are down-regulated compared to the wild bean (Bellucci et al. 2014). Thus, comparing the transcriptomes (gene expression level) of the crop and their wild relative species is necessary to decipher the genetic pathway modified transcriptionally during domestication. In tomato, studying

the domestication at a genome-scale level revealed the fixation of potential deleterious protein and expression level changes (Koenig et al. 2013). The selection of regulatory elements responsible for expression level changes, has been acting on clusters of co-expressed differentially expressed genes (DEGs) thus targeting pathways more than major effect genes (Sauvage et al. 2017). Following these findings, using a comparative analysis on potato and tomato, it was recently proved that the magnitude of domestication induced perturbation can expand to a complete pathway shutdown, such as the steroidal glycoalkaloids anti-nutritional pathway (Itkin et al. 2013).

Evolutionary, while comparing the regulatory changes to the domestication-associated genes under selection, in Maize, only one third of the DEGs were located on selected regions. They hypothesized that the remaining DEGs were *cis*-regulatory variants (variant acting on the gene expression of a linked gene, most probably transcription factor genes) that had "hitchhiked" along with the selected genes (Swanson-Wagner et al. 2012). The *cis*-regulations are often tissue- and stage-specific implying a strong impact on domestication-associated phenotypes when selected during domestication. The importance of these *cis*-regulatory variants highlights the necessity to study in depth the *cis*-regulatory elements in non-coding regions as well (e.g. promoter, intron, 5' untranslated region (UTR), etc.). Though, it appears that only the *cis*-regulatory regulation correlates with genes under positive selection due to domestication, while *trans*-regulatory elements do not (Lemmon et al. 2014).

### iii. Metabolome changes at the genome-wide scale

The study of the transcription of protein-coding genes reveal regulatory aspects of metabolic network behavior (Carrari et al. 2006). Another way to study domestication changes is to look directly at the gene expression products. The metabolome is considered as the bridge between the genome and the phenome. While selecting for phenotypes related to fruit taste and nutritional value, the selection directly targeted natural compounds also called secondary metabolites. In the wild, the important diversity of metabolites has a clear ecological role as to increase the potential of adaptability. Wild plants product and store many compounds, with rare biological activity, to be more adaptable in case of evolutionary challenges, the so called 'Screening Hypothesis' (Firn and Jones 2003). The variability in natural products is a source of potential protection (Lewinsohn and Gijzen 2009) against pathogens or against stresses due to climatic conditions (Langenheim 1994; Harborne 1999; Croteau et al. 2000; Gershenzon and Dudareva 2007), and simultaneously volatile organic compounds emitted by the flowers induce pollinator attractions (Pichersky and Gershenzon 2002).

During domestication the production of metabolites is shifted towards human interest. Many products are known to be selected in crops or counter-selected as previously mentioned with the shutdown of the glycoalkaloids anti-nutritional pathway (Itkin et al. 2013). Several studies used metabolomic data set to decipher the specific domestication related changes in secondary metabolites. In tomato for example, the crop experienced a loss of around 95% of the genetic and chemical diversity of its wild relative species *Solanum pennellii* (Perez-Fons et al. 2014). Recent metabolite based-GWAS studies performed on tomato confirmed the rewiring of the crop fruit metabolome during domestication (Sauvage et al. 2014; Zhu et al. 2018).

These comparative analyses of crop and wild metabolite diversity highlight the importance for modern breeding to improve crop nutritional quality and nutrient assimilation (Meyer et al. 2012b). The metabolomic changes that are due to domestication, namely the reduction in metabolite diversity, remains within the crop wild relative species. The modification of regulatory genes could enhance and elicit trait improvement (Harrigan et al. 2007).

Thus nucleotide, transcriptomic and metabolomic diversity are highly impacted by domestication. In this context, the use of crop wild relatives for crop improvement becomes necessary, and the use of “omics” technologies provides an opportunity to integrate and compare all levels from phenotype to genotype (Langridge and Fleury 2011).

### d. Crop wild relatives, the source of potential for improvement

The strong and constant human selection of domesticated phenotypes alters the selective pressures on cultivated plants and removes the process of natural selection. While shifting to a human environment, early domesticates experienced an increase of fitness for phenotypes with low fitness in the wild (Purugganan and Fuller 2009). In some cases, such selection induced the frequency decrease or even loss in the crop population of less desirable phenotypes (Zohary 2004). Domestication often selected against traits that increase plant’s defensive or reproductive success in natural environment which implies that domesticates became unable to survive outside the man-made environment, they lose their potential of adaptability (Gepts 2004; Pickersgill 2007; Allaby et al. 2008; Purugganan and Fuller 2011). The resulting relaxation of natural selection and creation of a human selective pressure induced genetic responses to domestication in domesticates, and this from early stages on (Zeder 2012; Marshall et al. 2014).

As previously mentioned, the domestication syndrome improved production related traits but simultaneously induced loss of fitness related to diseases and stresses resistances and/or tolerance. The modern breeding efforts were focused on modern cultivar diversity before to recognize the potential of remaining traits of interest within the crop wild relative (CWR) species genomes. These CWRs are the primary accessible source of diversity for crop improvement, therefore, there is an urge for conserving and studying their gene pool within their location of origin as well as in conserved seed stock.

Most of the past breeding improvement efforts were focused on increasing the yield and inducing resistance to pathogens. These breeding programs relied on core collections of cultivars, and since the very beginning the field looked into wild relatives to induce resistances that remained in the wild resources. In this context, few challenges of the next breeding era need to be tackled, one of them is the conservation of CWRs within their original wild locations but as well within conserved seed stock (Brozynska et al. 2015). The second challenge is to sequence and decipher the possible genetic, transcriptomic and metabolomic resources present within CWRs (Henry and Nevo 2014). The few comparative transcriptomic and genomic studies on crop and their CWRs revealed that, fortunately, most of the modified traits (ranging from drought tolerance to disease resistance) were preserved in the close relatives and can potentially be retrieved through introgression or genome editing (Eshed and Zamir 1995; Henry 2012). In the cultivated tomato the use of germplasm donor from wild tomato species such as *S. pennellii or S. habrochaites*, leaded to enhance the agronomic performance of *S. lycopersicum*, the cultivated form. So far, the use of CWRs has contributed for more than 50 improved traits notably related to disease resistances in the tomato species only but CWRs also increased quality and improved yield. The CWRs use to improve crops has been developed in many other species as wheat, rice, sunflower, potato notably (Brozynska et al. 2015). In this context, the scientific community extend the genomic researches from the crop to the crop wild relatives as well (See *II.a. The plant family: Solanaceae* section).

While traditional plant breeding was based on phenotypic selection, new methods such as forward genetics rely on the available and high-quality genomic data to detect the gene of interest location. The “omics” techniques allow the association of high-throughput genotyping and phenotyping to identify candidate genes. Therefore the current crop improvement focus mostly in improving modern cultivars with the genetic variation available in the germplasm collections (Langridge and Fleury 2011). In addition, the use of metabolomic-assisted breeding was as well proposed to improve the plant variance in metabolite composition (Fernie and Schauer 2009). The

crop wild relative species may be a source of novel alleles to improve the productivity, adaptation, quality and nutritional value of modern cultivars, as previously mentioned (Fernie et al. 2006).

A recent study took up the challenge of producing a *de novo* domesticated tomato, by editing the genome of a wild relative. Knowing the main genes involved in tomato domestication, the study reproduced the phenomenon of domestication with new genome editing methods (CRISPR-Cas9). They targeted 6 genes well known to be involved in domestication traits: *SP*, *OVATE*, *FAS* and *fw2-2*, *MULTIFLORA* (increasing the fruit number in crop) (and LYCOPENE BETA CYCLASE (*CycB*: increasing the content in lycopene thus the fruit nutritional quality). The study revealed the potential of mutating wild plants to reproduce domestication process (Zsögön et al. 2018). Not only such study offers the opportunity to study domestication more thoroughly but it also opens a new range of potential improvements by the retrieval of wild adaptability traits and the induction of traits of interest for crop production and human consumption.

In this context, it seems necessary to intensify the efforts to identify the changes that modified the wild progenitors into crop species. Indeed, by improving the comprehension of the past evolutionary process, it will help developing future strategies of breeding improvement.

### e. Potential of demographic inferences to decipher domestication

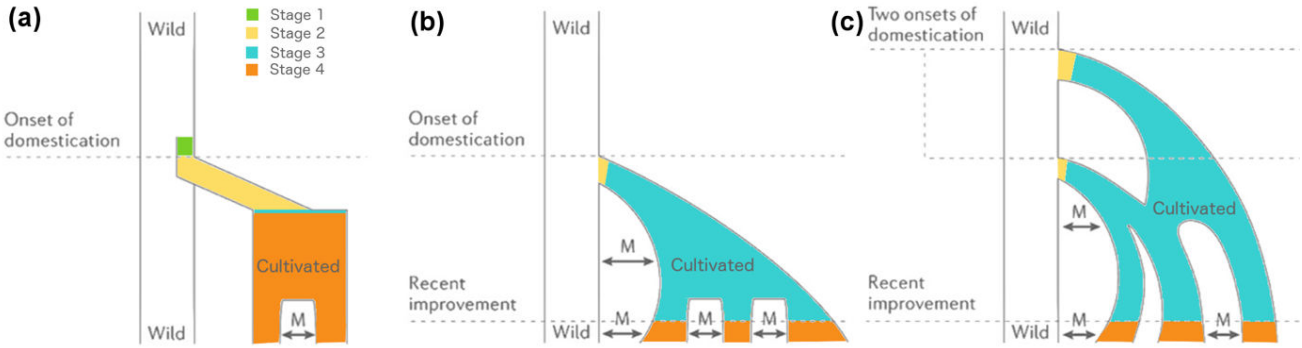



**Figure 5**. The characterization of domestication in crop species is dependent on understanding the initiation and the course of the domestication process. The width of the channels represents population size and geographical range; M = Ne*m, which is the product of effective population size (Ne) and the migration rate (m). (*a*) Earlier models of domestication posited a single domestication event and suggested that domestication occurred through strong selection and severe genetic bottlenecks in a small population of the wild progenitor, which resulted in greater reproductive isolation between the wild species and the domesticated species. (*b*) Alternative model including introgressions between cultivated and wild relatives. (*c*) Alternative model including introgressions and several onsets of domestication. *Adapted from: Meyer and Purugganan 2013*

Already Darwin had foreseen the potential of domestication for studying it as an artificial evolutionary process. Understanding the changes between crop and wild progenitors requires to decipher the evolutionary forces, namely the mechanisms of mutation, selection and drift. With the outbreak of unprecedented -omics data, it became possible to estimate the evolutionary processes by testing theoretical population genetic models, this approach is a model-based hypothesis testing. Using demographic models on genomic data offers a better resolution to estimate demographic parameters that impacted the stages of domestication (Cubry and Vigouroux 2018). Most current studies on domestication are highlighting the stage 2. This step of cultivation is paired with a strong filtering only on desired phenotypes, the resulting selective sweeps imprinting the crop genetic constitution by decreasing the whole genome diversity (Galtier et al. 2000). The reduction in nucleotide diversity is detectable by a correlated reduction in effective population size (figure 4b). With the combined use of genomic resources and demographic inferences between crop and wild progenitors, it becomes possible to detect and characterize these stages, through the changes in the effective population size and gene flow rate, and estimate their duration (Gaut et al. 2018).

The early demographic models of domestication suggested a single event of domestication with a severe bottleneck from a few individuals of the wild progenitor species (Haudry et al. 2007), resulting in a reproductive isolation between crop and wild especially in a case of re-localization of the crop to a human-environment (as illustrated in figure 5*a*). With the increase in genomic data and archeological records of crops, new general models have been proposed to fit the evolutionary histories of more crops. The genetic bottleneck that was supposed to be severe appeared to be variable according to the species, annual crops such as maize (Hufford et al. 2012) and tomato (Koenig et al. 2013) experiencing the expected and strong reduction in nucleotide diversity, but this reduction being minimal in perennial plants such as apple (Cornille et al. 2012). This underlines the influence of life history in the domestication scenarios, annual plants experiencing stronger selection generation after generation due to the inbreeding when perennials are obligate outcrossers and experience high rates of intra- and interspecific gene flow (Savolainen et al. 2007; Miller and Gross 2011). The recovery from a bottleneck can be highly improved by introgressions from wild relatives, therefore, a strong isolation is not necessarily a feature of domestication (Dempewolf et al. 2012) and the migration has to be included in domestication model (figure 5*b*). Several crops such as maize, pearl millet, wheat, carrot and tomato are thought to have experienced a single domestication event (Meyer et al. 2012a). But alternative models need to include the potential multiplicity of domestication events, as it is expected to characterize one quarter of the global food crops (Meyer

et al. 2012a). Some species such as barley, common bean or eggplant are expected to follow such patterns of domestication with parallel events in different regions or at different time points (figure 5*c*).

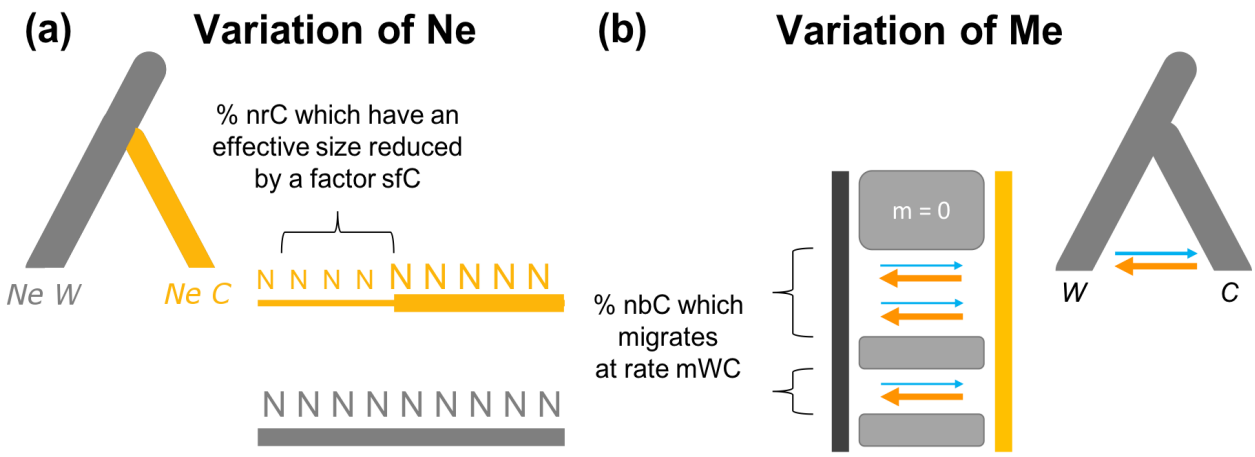



**Figure 6**. Representation of the heterogeneity across the genome. Briefly, *Ne* correspond to the effective population size (*Ne W*: wild, *Ne C*: crop), migration is shown by orange (from crop to wild) and blue (opposite direction) arrows. (*a*) Heterogeneity of effective population size, genome location that experienced a selection at linked neutral sites. The *%nrC* is the proportion of the genome that has been "selected", the extent of effective population size reduction in "selected" regions is *sfC* (in the crop). (*b*) Heterogeneity of the effective migration rates that highlights hotspots of introgression. The *%nb* is the proportion of "not barrier" regions (*nbW*: wild, *nbC*: crop). *Inspired by: Roux et al. 2013 & Sousa and Hey 2013*

The studies on proximity between crop and wild relatives of wheat and maize were based on phylogenetic tree based on distances at first (Heun et al. 1997; Matsuoka et al. 2002), but such method assumes that gene flow is negligible. The gene flow is not a limiting factor in model-based inferences, thus, using demographic inferences is better suited for the reconstruction of domestication processes (Cubry and Vigouroux 2018). The use of outgroup species defines ancestral or derived each allele at polymorphic sites. The frequency of polymorphisms shared between both is summarized in a joint site frequency spectrum (jSFS). From this summary statistics of the population genetic diversity, the inferences can be made using likelihood or pseudo likelihood approaches such as proposed in the software FastSimCoal (Excoffier et al. 2013), ðaði (Gutenkunst et al. 2009) or Approximate Bayesian Computation (ABC). Bayesian computation is based on extensive simulation of potentially complex models and on assessing if the model fits the observed data. Though on large data sets such as RNAseq data, the coalescence method is faster and efficient on two populations (e.g. crop and wild) using a jSFS with given ancestral states (Marin et al. 2012). The coalescence

method implemented in ðaði infers most probable scenarios of domestication and rebuilds population history with a diffusion-based approach. Such approach was used to investigate and decipher the Asian rice domestication, for example (Molina et al. 2011).

The different models of divergence can implement scenarios of strict isolation (figure 5*a*), of isolation with migration (figure 5*b*) or of secondary contact (Gutenkunst et al. 2009). The variation in effective population size across the genome (figure 6*a*), such as a local reduction of *Ne*, due to a selection at linked neutral sites, known as the Hill-Robertson effect (Hill and Robertson 1968), is implemented by considering two categories of loci (Sousa and Hey 2013). The identification of potential genomic hotspots of introgression (figure 6*b*) is implemented, as well, by clustering the loci in two categories with different effective migration rates (Roux et al. 2013). While most demographic studies reconstruct the domestication events in a single species (Eyre-Walker et al. 1998; Wright, 2005; Sabeti et al. 2007; Zhu et al. 2007), the question of the convergence between the domestication processes among different crop and wild species of a same family received little attention so far. Ðaði has been recently used in order to decipher the most probable domestication history of the rice comparing wild and crop species (Molina et al. 2011) and allowing the estimation of splitting time between ancestral populations, number of migrants per generation between the populations and estimating the strength of the bottleneck. This study and some others on humans (Lindblad-Toh et al. 2005) and animals such as horses (Wade et al. 2009) or dogs (Ostrander et al. 2017) are evidencing that comparative analyses on the demography of several species would highlight the convergence of demographic modifications due to human history and animal and plant domestication.

The main interest of demographic inferences is to better understand the process of domestication. It brings valuable knowledge for modern breeding to understand the crop demographic changes coupled with the responses to this artificial selection process (Hammer 1984; Vigne 2011), but to human history and sociology, as well, by estimating dating for the transition to crop cultivation that correspond to the settlement of human populations (Zeder 2015). In this context, we used the common selected phenotypes in three domesticated species of the Solanaceae family to evaluate the extent of convergence between independent domestication events.

# II. Focus on the study systems: *Solanaceae*

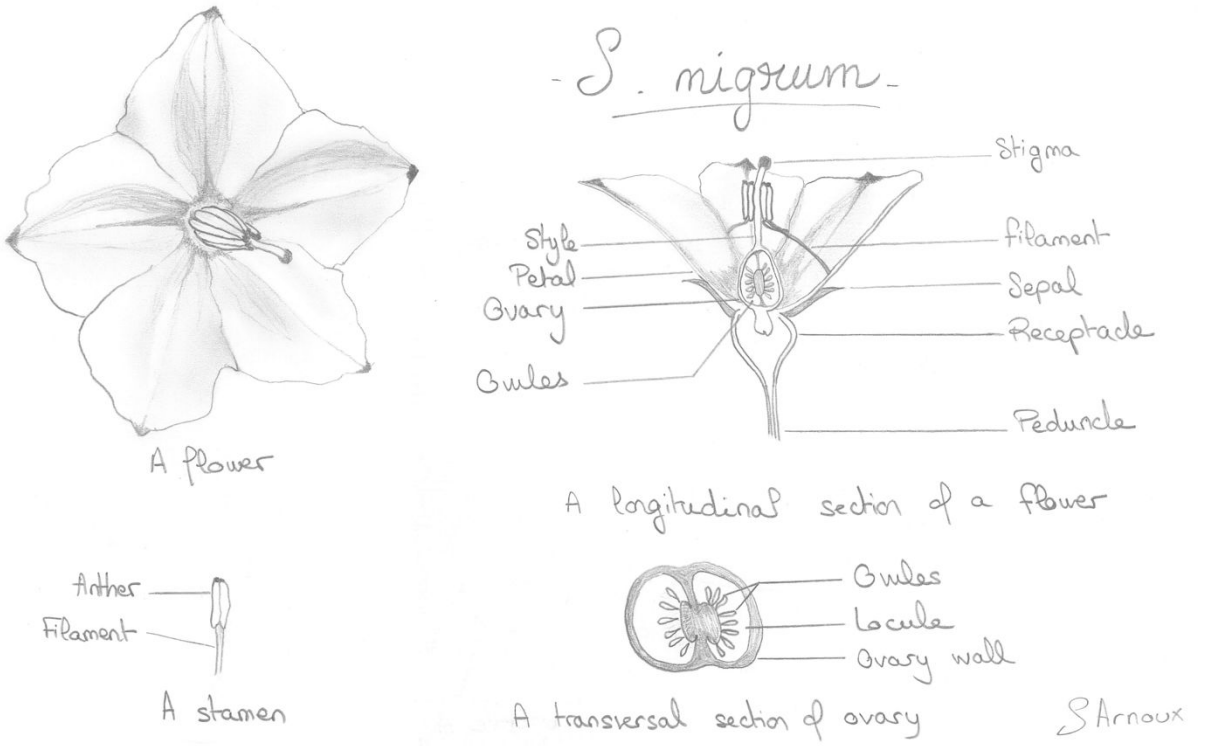



**Figure 7**. Botanic illustration of the Solanaceae family with an example in the *Solanum nigrum*.

## a. The Solanaceae family

The Solanaceae is the most important angiosperm family in terms of species number. The Solanaceae or nightshade family includes ~3,000 species distributed in 90 genera (Vorontsova and Knapp 2012). Both genetic and species level diversity in the family is mainly concentrated in the Andes of South America, and the family has a classic Gondwanan origin explaining the worldwide distribution of its species. The largest genera is the *Solanum* L. with around 1,500 species (figure 7) (including three of the most important crops: the cultivated potato (*Solanum tuberosum* L.), tomato (*S. lycopersicum*) and eggplant (*S. melongena*). The second genera with relevant importance is the *Capsicum* genera, composed of 30 species, that includes five domesticated species of pepper (*C. annuum* L., *C. frutescens* L., *C. chinense* Jacq., *C. baccatum* L. and *C. pubescens* Ruiz et Pav.). To discuss species within the Solanaceae family, it is important to precise that species, especially in the plant

kingdom, are defined according to common phenotypic and now genetic traits, but that reproductive barrier is not always present. Two separate plant species can still interbreed and have fertile progenies within hybrid zones (Barton and Hewitt 1989). Regarding domestication, most of the crop species can still interbreed with their wild relative species.

### i. Economic importance

This taxonomic family includes a number of commonly collected or cultivated species within which several species are leader in the economic and production fields. These species are represented on the figure 8. Although, the food production per year is dominated by cereals, because of their high nutritional values, the Solanaceae family comes second. In comparison, cereals represent around 50% of the global production, with maize (1,038 Mt), rice (742 Mt) and wheat (733 Mt), and Solanaceae follows with potato (381 Mt), tomato (172 Mt), eggplant (50 Mt), pepper (35.7 Mt) and for more economic reason tobacco (7.2Mt). When considering the economic importance for export, the Solanaceae is one of the leaders in export, just after cereals with tomatoes (9.1 billion US$ of export value), tobacco products (5.7 billion US$), potatoes (4.8 billion US$), chilies and peppers (1.4 billion US$) and eggplants (0.45 billion US$). The detailed economic and production values are given in table 1. This thesis work focuses on eggplant, pepper and tomato that are three of the five most important economical Solanaceae.

**Table1**. Main crops export and production values for the year 2014. *Source: FAO 2014*

| Species | Export value (billion US$) | Production value (Mega-tons (Mt)) |
|---|---|---|
| Wheat | 47.7 | 733 |
| Maize | 32.8 | 1038 |
| Rice - total (Rice milled equivalent) | 26.0 | 742 |
| Tomatoes | 9.1 | 172 |
| Chillies and pepper (green and dry) | 6.2 | 35.7 |
| Tobacco products | 5.7 | 7.2 |
| Potatoes | 4.1 | 381 |
| Eggplants (aubergines) | 0.5 | 50 |

### ii. Scientific recognition

An important scientific community addressed the questions of plant adaptation and diversification in the Solanaceae with the International Solanaceae Initiative (SOL) that has a long-term goal of creating a network of resources and information about Solanaceae genomes. Within the SOL network, an effort was made to produce a clade-oriented database dedicated to the Solanaceae, namely the Solanaceae Genomics Network (website at https://solgenomics.sgn.cornell.edu/). Therefore, considerable genetic resources of natural diversity in tomato (*S. lycopersicum*), eggplant (*S. melongena*) and pepper (*Capsicum spp*) are available and will constitute the raw material of this study. Part of these genetic resources have been characterized at the phenotypic and molecular levels while core collections have been constituted to investigate the genetic architecture of traits of agronomical interests through various approaches (QTL mapping, GWA). The three species, that are at the root of the project, are of major scientific interest especially with the tomato being one of the first genetic model for genetic diversity study or fleshy fruit development.

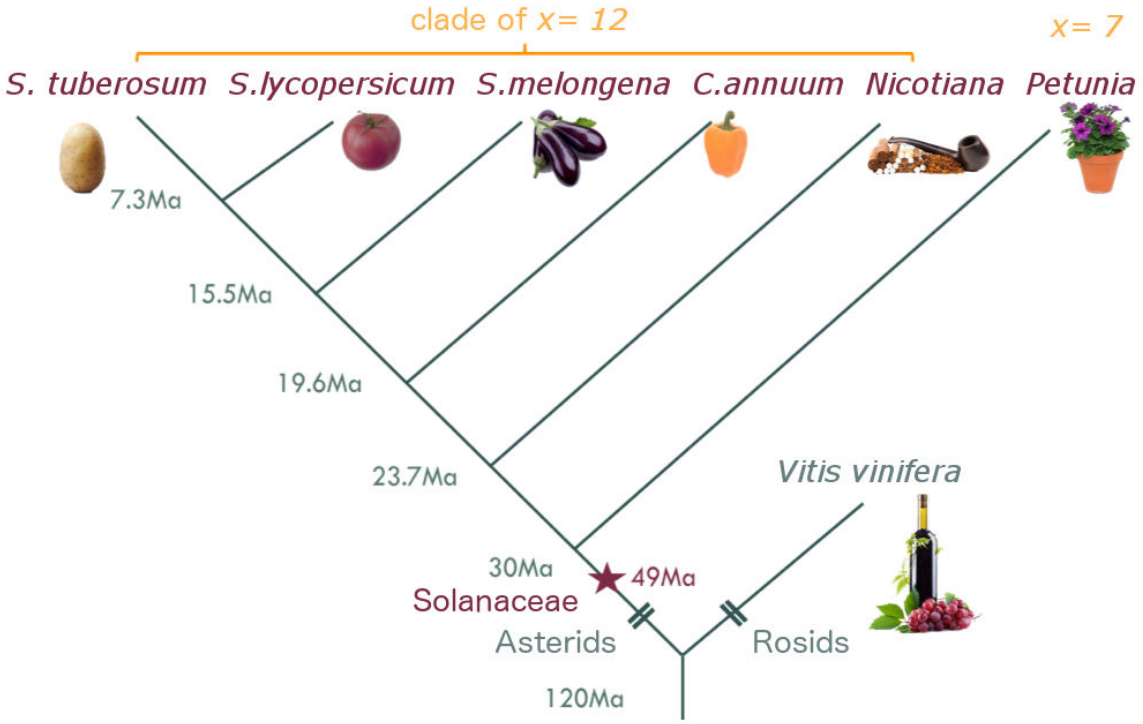



**Figure 8**. Phylogeny of the Solanaceae family, showing the family-specific Solanaceae hexaploidy event shared with most eudicots. Solanaceae mutation is placed before the divergence of Petunia and the x = 12 crown-group (~30 and 49 Million years ago (Ma)). *Inspired by: Bombarely et al. 2016*

Most of the taxa of both the *Solanum* and the *Capsicum* genus have a chromosome number of *n* = 12 (figure 8). The three species have highly colinear and syntenic reference genome maps available (Wang et al. 2008), that facilitates the comparative genomic analyses. The *S. lycopersicum* reference genome (The Tomato Genome Consortium 2012) is regularly updated (currently at version 3.2), the *C. annuum* reference genome (Qin et al. 2014) has, as well, a version 2.0 and the new reference genome of *S. melongena*, yet unpublished (Lanteri et al. 2014; The Eggplant Genome Consortium 2017), was obtained in December 2017 (Table 2) thanks to a project coordinated by the University of Torino, Italy, who gave us a private access to the complete genome sequence. Moreover, the 3 species have different geographical origins, tomato and pepper originating from south and central America and eggplant from Asia. These independent histories are a great opportunity to test if the domestication outcomes are parallel, convergent or species-specific within a same plant family. Indeed, despite their independent geographical and domestication histories, these three species experienced a strong selection pressure on common phenotypes such as fleshy fruits (The Tomato Genome Consortium 2012) and other traits that added interest for human culture and consumption.

**Table 2.** Genome feature of the three species eggplant, pepper and tomato. *Source: Arumuganathan and Earle 1991; The Tomato Genome Consortium 2012; Lanteri et al. 2014; Qin et al. 2014*

| Genome Features | *S. melongena* | *C. annuum* | *S. lycopersicum* |
|---|---|---|---|
| Mating system | | self-compatible | |
| Assembled genome size (Gb) | 1.2 | 3.349 | 0.76 |
| Accession - Version | *(67/3) - NA* | (Zunla-1) - v.2.0 | (Heinz 1706) - v.3.2 |
| Number of scaffolds | 10,383 | 967,017 | NA |
| Contig N50 (bp) | 1,060,000 | 1,226,833 | NA |
| GC content (%) | 35.7* | 34.9 | 34.0 |
| LTR rate (%) | NA | 70.3 | 50.3 |
| Predicted protein-coding genes | 34,916 | 35,336 | 34,769 |
| Gene number used in our analyses | 18,047 | 19,628 | 17,545 |
| Average gene length (bp) | NA | 3,363 | 3,006 |
| Average CDS length (bp) | NA | 1,020 | 1,063 |
| Sequence anchored on chromosome (%) | 68.79 | 78.95 | 98 |
| Genes anchored on chromosome (%) | 81.43 | 88.29 | NA |

*Data recovered from the draft reference genome of eggplant. *Source: Hirakawa et al. 2014*

In the following sections, the three species (i.e. eggplant, pepper and tomato), will be described thoroughly. Though, it is important to precise that the three species have a high strong level of synteny between their respective genome, which allows and facilitate eventual transversal applied and basic research (e.g. improvement of a crop species with knowledge on syntenic genes within another crop species, study of the evolution process of the Solanaceae family). Already some resistance genes were found to analogs and present in the two or three of these species, namely the Sm7RGA4 in *Solanum melongena* and *C. annuum*, or Sm7RGA8 in *S. lycopersicum* and *C. annuum* (Reddy et al. 2015). The close genetic structure or collinearity allows a real comparative analysis of the three species and involve potential trans-specific breeding improvement between the species. While performing mapping of the eggplant or the pepper (see following sections), the synteny was as well intensively used to facilitate the genomic architecture understanding (Doganlar et al. 2002a; Hirakawa et al. 2014). Already, while studying eggplant, Doganlar highlighted that 40% of the QTLs had orthologs in at least one of the species including tomato, potato and pepper (Doganlar et al. 2002b). The genome evolution of Solanaceae was highly impacted by domestication and the crop species were studied to identify genes for domestication (Doganlar et al. 2002b) and morphological changes (Frary et al. 2002). These studies were pioner of crop comparative analyses, and intitiated a better understanding of the genome changes due to domestication in Solanaceae family such as for genetic and molecular regulation of domestication-related traits in tomato and pepper (Paran and Van Der Knaap 2007).

## a. Eggplant history

### i. Taxonomy and species history

Eggplant (*Solanum melongena* L.) is a vegetable originally growing in warm-weather conditions such as in tropical or subtropical regions (figure 9*A*). The cultivated eggplant has a large, oblong, purple-skinned “Black Beauty”-type fruit, but wild and semi-domesticated eggplant usually present a small, round, yellow fruit and a plant with abundant prickles (figure 9*B*). *S. melongena* and its wild relatives are part of the clade of “spiny solanums” named Leptostemonum (Levin et al. 2006), the clade is originating from Africa where wild species are still present (Knapp et al. 2013; Meyer et al. 2015). Wild species of eggplant moved to tropical Asia with step-wise expansion where *Solanum melongena* L. was domesticated and back to the Middle East as feral forms (Weese and Bohs 2010; Meyer et al. 2012a). The eggplant domestication remains debated and several parallel events of

domestication were proposed according to writing records that recall its presence simultaneously in China and south eastern Asia around 2,000 years ago (Suśruta and Bhishagratna 1907; Meyer et al. 2012b).

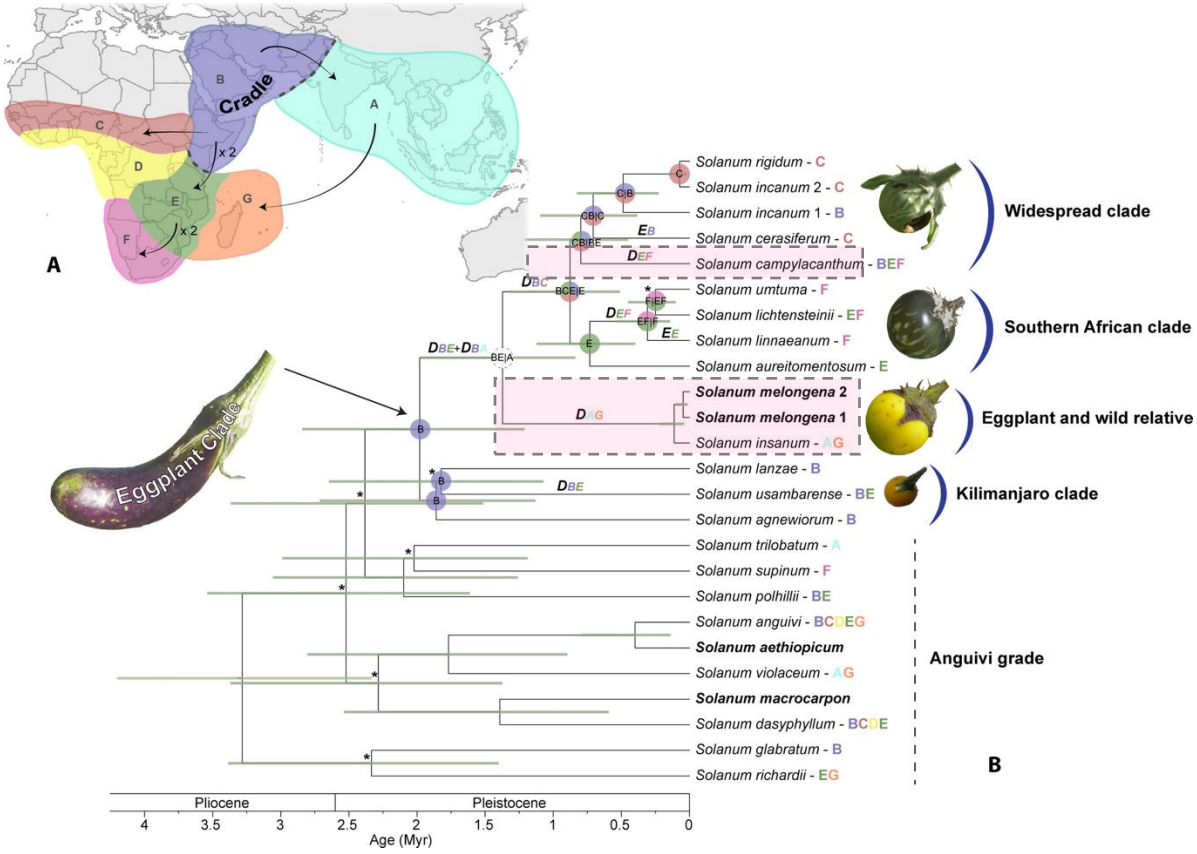



**Figure 9**. Phylogeny and biogeography of the Eggplant clade based on whole chloroplast genome sequences. The pink boxes highlight the species used within this Thesis work. (A) Map showing the seven biogeographic areas used to infer the biogeographic history of the Eggplant clade. (B) Full-plastome dated phylogeny of the Eggplant clade. The most probable ancestral area is figured at each node of the Eggplant clade; high levels of biogeographic uncertainty are indicated with dotted lines. *Source: Aubriot et al. 2018*

The first taxonomic studies on wild accessions were performed according to morphological traits but were insufficient to classify all the species (Lester 1986). The first phylogenetic analyses of wild relatives on chloroplast DNA (cpDNA) was performed on nine species including: *S. aethiopicum*, *S. anguivi*, *S. gilo*, *S. incanum*, *S. integrifolium*, *S. macrocarpon*, *S. olivare* and *S. panduriforme* (Sakata

et al. 1991). Following this study, the phenetic method was used to provide a cladistic taxonomy of 36 accessions of crop and wild relatives forms of series *Inaciformia, Macrocarpa* and *Aculeastrum* (Mace et al. 1999). It is with RFLP analysis of the mitochondrial DNA that six related species of *S. melongena*, namely *S. gilo*, *S. integrifolium*, *S. indicum*, *S. sanitwongsei*, *S. surattense* and *S. torvum* were phylogenetically classified in 2003 (Isshiki et al. 2003). Though, using compiled details from AFLP and morphological traits, a cladistic method suggested that the taxonomy of the *Solanum* sections and subgenera including several species had to be reconsidered (Furini and Wunder 2004). Eggplant taxonomy was challenging for long but a recent study compiled 2 nuclear and 3 plastid regions, and used a phylogenetic tree based on a maximum likelihood and a Bayesian inference, that included 42 of the 56 recognized species to decipher the entire clade taxonomy (Aubriot et al. 2016).

The crop species experienced several taxonomic discussions, it was structured in 3 morphoforms (group E, G, H), within a study on crop and wild relatives including , the study used cpDNA for phenetic and cladistic methods (Sakata and Lester 1997). The group were considered as artificial and the wild relative progenitor was hypothesized in a study on *S. melongena*, *S. incanum* and *S. insanum* from Karihaloo in 1995 (Karihaloo and Gottlieb 1995). This wild progenitor was recently ascertained, after being long debated, in a review of taxonomy from 2016 (Ranil et al. 2016). Both species remain in sympatry within Asia.

The genetic diversity of eggplant cultivars is reduced compare to their wild progenitor. The capacity for wild and crop eggplants to hybridize producing fertile plants increase the potential use of crop wild relative to improve modern cultivars (Davidar et al. 2015).

### ii. Genetic resources

The eggplant ranks in the top five important vegetables and is very important in Asian regions. The yield of eggplant is really dependent on climatic conditions (Frary et al. 2007), therefore, it is important to improve modern cultivars to face global warning effects by identifying and using CWR diversity. As for many crops, only few eggplant cultivars are cultivated worldwide (Muñeoz-Falcón et al. 2009) and consequently the varietal diversity is mainly concentrated within the original locations that retain a diversity in number of cultivars (Ali et al. 2011). In southeast Asia and India, thousands of local landraces exist and represent a wide range of variability in morphology, flavor and pathogen resistance. Wild relatives were used in breeding to induce resistances such as for the bacterial wilt resistance present in the wild species *S. macrocarpon*, *S. gilo* and *S. viarum* (Reddy et al. 2015). But in 2016, Syfert et al. related the absence of modern cultivars with introgressed traits from wild

relatives (Syfert et al. 2016), the breeding improvement relying mostly on genetically modified cultivars such as the *Bt* eggplant (Bhagirath and Kadambini 2009). Following these technological advances, 23 populations of wild relatives were studied to identify the potential crop-to-wild gene flow. The wild eggplant requires a pollinator visit for the pollen to transfer, in result the wild are 5-fold more outcrossing than the domesticated, and the study highlighted the capacity of hybridization that would lead to introgression from crop to wild genepool a possible concern if the domesticated were genetically modified (Davidar et al. 2015).

Despite the little use of crop wild relatives in modern cultivars, the biodiversity of wild relatives of eggplant remains in the landraces and in the wild relatives. Thus a large-scale effort was made by the scientific community to collect and conserve germplasm of wild relatives and landraces encompassing more than 15,000 accessions in 99 institutions worldwide for landraces alone (Meyer et al. 2012b). In an effort to produce a public database, the European Database for Eggplant was developed within the framework EGGplant genetic resources NETwork, the platform offered three independent search pages, on databases of eggplant, Solanaceae, and on Solanaceae bibliography (http://www.bgard.science.ru.nl/WWW-IPGRI/eggplant.htm).

### iii. Molecular markers and genome mapping

This work of genetic resources collection, was complemented by the production of molecular markers to develop a linkage map of the eggplant reference genome. The first one, named eggplant-LXM 2002, was based on 58 F2 individuals from an interspecific cross between *S. linnaeanum* (MM195) and *S. melongena* (MM738). This map provided 233 RFLP markers that were used to decipher the synteny analysis of the genomes of eggplant, tomato, potato and pepper (Doganlar et al. 2002b, a; Frary et al. 2003).

In parallel, another linkage map based on 88 RAPD and 93 AFLP markers, at first, was complemented over time, by 236 SSR markers and spanned a total genetic distance of 959.1 centimorgans (cM) in 14 linkage groups. This map aimed to facilitate breeding programs by mapping fruit shape and color development traits (Nunome et al. 2001, 2003, 2009).

Another map, the eggplant-COSII map was produced in GAFL-INRA by Marie-Christine Daunay and was based on 58 F2 individuals from an interspecific cross between the same accession as the eggplant-LXM 2002. Using the tomato synteny, they provided 232 markers COSII of the eggplant genome (Wu et al. 2006, 2009b).

Following these two linkage maps, the production of a draft reference genome was published in 2014 composed of 33,873 scaffolds termed SME_r2.5.1 that covered 833.1 Mb of the eggplant genome (~74 %). They identified 56 conserved synteny block between tomato and eggplant (Hirakawa et al. 2014). Recently, a genome (yet unpublished; conference citation: The Eggplant Genome Consortium 2017) was produced within the Eggplant Genome Consortium, this reference genome of *S. melongena* L. inbred line '67/3' was sequenced with a combination of Illumina sequencing and optical mapping, and covered 1.06 Gb of the 1.2 Gb eggplant genome, anchoring 78.79% of the sequences produced in the final assembly, more details are provided in the table 2.

## b. Pepper history

### i. Taxonomy and species history

The first taxonomic studies on *Capsicum* was produced in a monography by Fingerhuth in 1832 (Fingerhuth 1832), where he depicted a detailed list of species from the generis Capsici including the *C. annuum* L.*, C. frutescens* Willd., *C. baccatum* L., *C. microcarpum* D., *C. sinense* Jacq. that would be later on denominated *C. chinense* Jacq.(Heiser and Pickersgill 1969). It is only in 1953 that four cultivated species were officially recognized with the first description of the cultivated peppers, namely *C. pubescens*, *C. annuum*, *C. baccatum* (called *C. pendulum* in the study) and *C. frutescens* (Heiser and Smith 1953). The fifth cultivated species *C. chinense* (called *C. sinense* in the study) was included few years later (Smith and Heiser 1957). The use of allozyme of domesticated and wild taxa of *Capsicum* helped deciphering the genera. Despite the easily discernible white-flowered and purple-flower group, the results showed that discerning the species within groups is problematic. They highlighted the similarities in the species *C. baccatum* and the *C. praetermissum*, resolving they were part of the same species, and they named the *C. annuum* complex when they could not disentangle the species *C. annuum v. annuum*, *C. chinense*, and C. *frutescens* (Jensen et al. 1979).

Within this taxon, the "bell pepper" (*Capsicum annuum* L.) is a vegetable originally growing in the warm-weather conditions of tropical Mesoamerica) (figure 10A). *C. annuum* is the most widely grown spice and is worldwide bred and consumed. The only distinguishable trait that distinguishes *C. annuum* from its wild relative species is the rate of germination. Traits relative to the domestication syndrome such as fruit size, position and loss of seed shattering vary among landraces) (figure 10B). The archeological records are too limited to detect which traits arose first from fruit shapes, color

and degree of pungency (Loaiza-Figueroa et al. 1989). Most of the remains (mainly seeds) were located in caves in the Tehuacán valley in Mexico but they are within the same size range of modern crop wild relative *C. annuum* var. *glabriusculum* (Kraft et al. 2014). The wild progenitor of the current cultivated pepper was part of the human diet since about 9,500 years Before Common Era (BCE). The pepper is one the only crop for which farming people still consume wild species as much as their cultivated descendants. The importance of Pepper comes from its pungency that is used as spices. The unique archaeobotanical record identified, as certain *C. annuum*, was estimated to be old of 1,500 years BCE (Lentz et al. 1996).

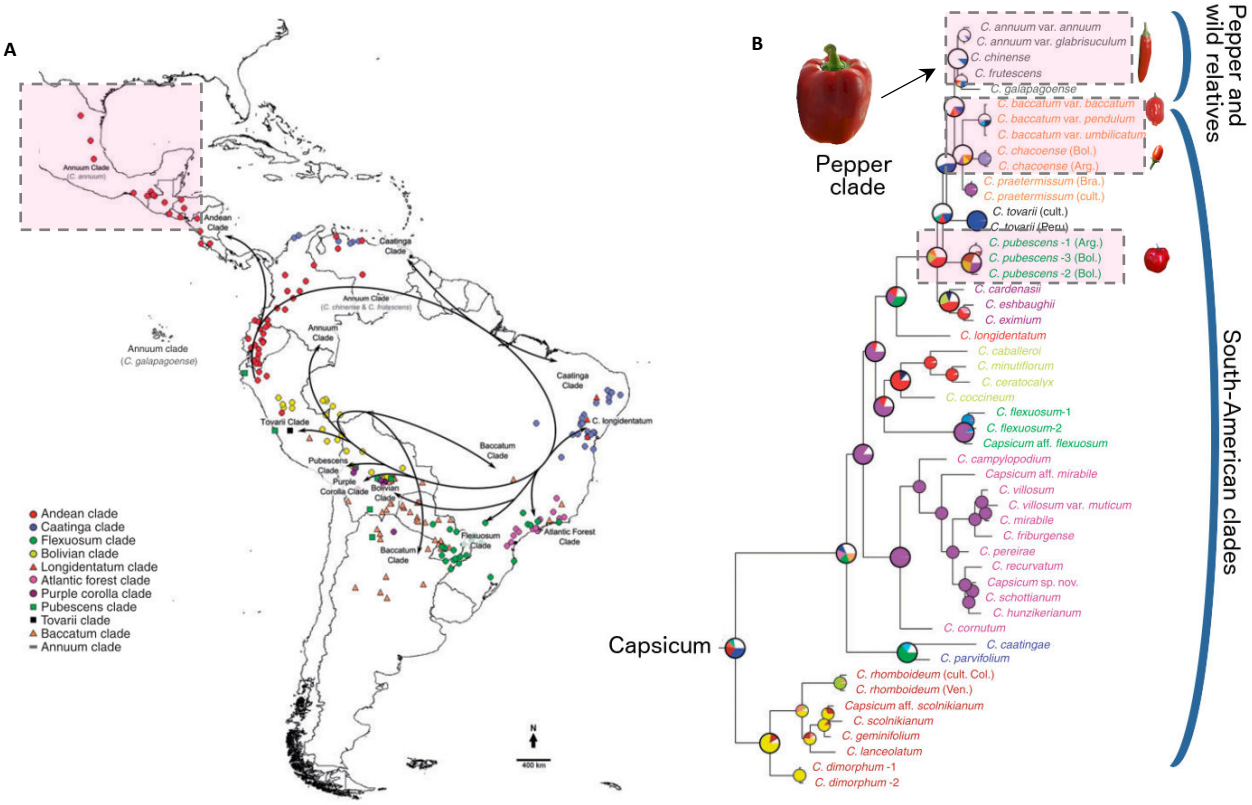



**Figure 10.** Hypothesis of Capsicum expansion. The pink boxes highlight the species used within this Thesis work. (A) Schematic expansion of the species. The arrows represent clades and monotypic lineages going across and/or pointing to the areas inhabited by their species. (B) Ancestral areas reconstructed by Bayesian MCMC analysis. Pie charts are larger for the main nodes to make them more evident. Color codes reflect the major clades based on the phylogenetic results (grey scale for the Annuum Clade). Markings in different colors/shapes indicate selected population localities. *Source: Carrizo García et al. 2016*

A recent study combining ecology, archeology, linguistics and genetics of *C. annuum* identified a potential domestication center in central-east Mexico and suggested a timing estimation of 6,500 years BCE for the domestication supported by few other archeological records and timing elements of the proto-Otomanguean language (Kraft et al. 2014). If the hypothesis concerning the starting time of pepper domestication remains uncertain, the cultivation and export of wild pepper species around America by the ancestors of native people, led to five independent domestication events (the original location of the *Capsicum* species are represented on the figure 10A). Thus, *C. annuum* was initially domesticated in central-east Mexico but *C. frutescens* was in the Caribbean, *C. baccatum* in lowland Bolivia, *C. chinense* in northern lowland Amazonia and *C. pubescens* in the mid-elevation southern Andes (Eshbaugh 1983). These other cultivated species, such as the complex of species of *C. chinense* and *C. frutescens* considered as similar species (Pickersgill 1971; Walsh and Hoot 2001; Guzmán et al. 2005), were not drastically domesticated like *C. annuum* was, and, they are often used as improvement material for the cultivated pepper (Hill et al. 2013).

### ii. Genetic resources

Despite the five domesticated species of *Capsicum,* the modern breeding programs focused mostly on the non-pungent cultivars of *C. annuum* (Pickersgill 1997). A first comparative study between domesticated and wild *Capsicum*, using isozyme-coding loci, revealed a reduction of the total genetic diversity in the crop accessions (Loaiza-Figueroa et al. 1989). A following study focusing only on *C. annum* and using RFLPs, highlighted the lower genetic diversity in modern cultivars of “bell pepper” (non-pungent pepper) cultivars in Europe and North-America compare to the small-fruited accessions cultivated world-wide (Lefebvre et al. 1993). Both studies confirmed the expectation of the species using predominantly inbreeding as mating system, indeed almost all the species of *Capsicum* happened to be self-compatible (exception being *C. cardenasii*). The *Capsicum* domesticated species were reported to have low level of heterozygosity compared to the wild (Ibiza et al. 2012). From the first results, it became clear that genetic diversity would be collected more efficiently while favoring an extensive sampling of multiple populations (Brown and Marshall 1995). Thus, the crop wild relatives of pepper are important and constitute the source for further genetic studies and breeding improvement. They represent a gene reservoir that can bring solutions for agricultural problems such as conferring disease resistance of increasing quality and yield. In this

context, in 2001 they started testing 13 populations of *C. annuum* from Mexico to test for viruses resistances(Hernández-Verdugo et al. 2001).

The effort to collect core collection kept increasing while the genomic methods improved using the diversity as source of power. Indeed, 43 accessions of four species of cultivated pepper were characterized using 30K unigene pepper GeneChip revealing the genetic structure of the species (Hill et al. 2013). Following this, 1,352 non redundant accessions from 11 *Capsicum* species were genotyped using 28 microsatellites (SSR) to decipher the genetic diversity and structure of the genera (Nicolaï et al. 2013). They could show the clustering of each domesticated species but a strong discrepancy of the close wild relative, namely *C. annuum* var. *glabriusculum*) often referred as 'chiltepin', supposedly wild progenitor of *C. annuum*, but from these results apparently subdivided in species respectively progenitor of all domesticated species. Following these results, a core-collection of 332 accessions was established and maintained in INRA - CRB-lég (https://www6.paca.inra.fr/gafl_eng/Vegetables-GRC). This collection is completed by a germplasm bank of Zaragoza in Spain that contains 51 landrace accessions and 51 accessions from the complex of 9 species (González-Pérez et al. 2014). One-third of the world's pepper production is from China, thus in 2016, they contributed with 372 GenBank pepper accessions of Chinese local cultivars and landraces (Zhang et al. 2016b). In 2015, a study used pepper to show how gene bank could be improved selecting the accessions on the basis of diversity instead of selecting for specific traits (Van Zonneveld et al. 2015).

In the following work, the comparative analyses will focus on 4 species. The crop population used in our analyses is *C. annuum* that is the most cultivated domesticated form of pepper. In the demographic inference analyses we used the wild progenitor *C. annuum* var. *glabriusculum* to decipher the domestication process, and for the transcriptomic analysis we used *C. frutescens* and *C. chinense,* a complex commonly considered as same species, they both share the same location in the lower Andean and have a recurrent gene. This complex of species is considered and was already used as potential source of diversity for *C. annuum* improvement to resistance to diseases (Polston et al. 2006; Ibiza et al. 2010), pests (Fery and Thies 1997) and nutritional quality (Zewdie and Bosland 2000).

### iii. Molecular markers and genome mapping

In parallel to this work of genetic resources collection, molecular markers were used to develop a linkage map and a genome mapping. It is using RFLP that Prince et al. (1993) started the

linkage mapping using 192 molecular markers for Capsicum and the synteny comparison with the tomato. Following this, an effort was made to better understand the genome of *C. annuum* with mostly anonymous markers as RFLPs, AFLPs and SSR (Lefebvre et al. 1993; Prince et al. 1993; Paran et al. 1998; E. Z. Kochieva 2003; Adetula 2006; Akbar et al. 2010).

A first complete linkage map was proposed in 2006, this map comprises 381 markers including 271 Conserved Ortholog Set (COSII) using the synteny between pepper and tomato to position the markers in the pepper genome. The Pepper-COSII map was based on 94 F2 individuals from an interspecific cross between *C. frutescens* var. BG 2814-6 and *C. annuum* cv. NuMex RNaky. It was the first map representing the 12 contiguous linkage group corresponding to the respective chromosomes of the pepper genome including crop and related *Capsicum* species and spanning 1,613cM (Wu et al. 2006, 2009). In parallel, two maps were produced by private company, the Pepper-AC99 and the Pepper-FAO3 available on the Sol Genomics Network website (https://solgenomics.net/). Respectively, the Pepper-AC99 map was based on 100 F2 individuals from the inter-specific cross of *C. annuum* cv. NuMex RNaky and *C. chinense* var. PI159234, including 426 markers used to construct a linkage map of 1,304.8 cM.

The second, the Pepper FAO4 map was based on 100 F2 individuals from the cross of the *C. annuum* cv. NuMex RNaky and *C. frutescens* BG 2814-6, including 728 molecular markers and covering 1,358.7 cM of the pepper genome.

Following this mapping, the effort was pursued to improve the genome mapping and two reference genomes were proposed in 2014. An international group including scientists from Korea, Israel and USA presented the sequence of the hot pepper *C. annuum* cv. CM334 (Criollo de Morelos 334) with a 186.6x coverage using Illumina technology (Kim et al. 2014).

In parallel, scientist from China and Mexico published the complete genome of two *Capsicum* accessions, one Chinese cultivated Zunla-1 and one Mexican wild Chiltepin (Qin et al. 2014). The previous Zunla-1 reference genome is the one we used in our study and details are available in the Table 2.

Recently, in 2018, a linked-read sequencing technology was used to anchor over 83% of the final assembly, producing a high-quality reference genome (Hulse-Kemp et al. 2018).

## c. Tomato history

The cultivated tomato, *Solanum lycopersicum* L., is one of the most important crops from the Solanaceae family. It is a model organism with high economic and scientific value. The chapter I of this thesis describes into details the phylogeny, the taxonomy and the scientific history of population genomics in tomato, therefore, here, I will not extend this section. Briefly, the tomato was domesticated from its wild progenitor *S. pimpinellifolium* in Peru (figure 11*A* – estimated silent divergence of 0.6% - TGC, 2012) before experiencing two bottlenecks: first moving to Mesoamerica (Blanca et al. 2012) and then with few cultivars introduced to Europe from Mexico (Atherton and Harris 1986; Blanca et al. 2015). These events led to specific footprints with domestication and improvements sweeps (Lin et al. 2014). In the following work, the comparative analyses will focus on three species, the crop *S. lycopersicum*, the wild progenitor *S. pimpinellifolium* and the wild relative group peruvianum (figure 11*B*).

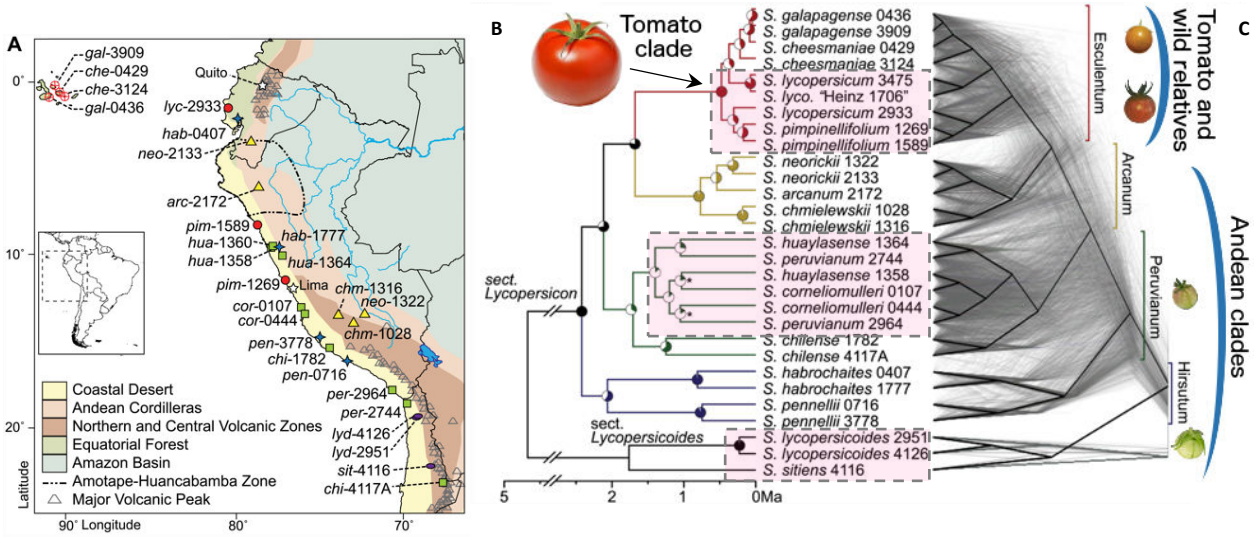



**Figure 11**. (A) Wild tomato species originally inhabit diverse ecological zones (shaded regions) along the western coast of South America and the Galápagos Islands. (B) A whole-transcriptome concatenated molecular clock phylogeny with section Lycopersicoides as the outgroup. Branch colors indicate the four major subgroups (labels on right). The pink boxes highlight the species used within this Thesis work. (C) A "cloudogram" of 2,745 trees (grey) inferred from nonoverlapping 100-kb genomic windows. For contrast, the consensus phylogeny is shown in black. *Source: Pease et al. 2016*

# III. Scientific questions and hypothesis of the thesis

Plant domestication considerably altered the modern cultivars used in current food production. The process of crops domestication is recurrently studied to answer the main questions concerning crop history:

- What was the wild progenitor species of the current crop?
- Where and when occurred domestication?
- How much domestication impacted genomes and transcriptomes of crop species?
- What were the genes and pathways targeted by selection?
- And finally, what can be retrieved from the wild relative species to improve modern cultivars?

My research project aims to answer some of these questions by revealing the extended footprints of domestication on the demographic history and on the expressed genes (is there any difference between gene diversity and expression profiles) of a trio of Solanaceae species: the eggplant, the pepper and the tomato. These three species have undergone independent domestication events and the wild population samples collected open a gate to study their genetic diversity and their phylogenetic history. The comparative study of these three species of Solanaceae is necessary to underlie the process of Solanaceae domestication. Indeed, by performing comparative transcriptomics, the description of matches and differences between crops and wild species allows to establish the domestication-associated footprints.

In the first chapter, the state of the art of research on tomato as model species gives an overview on the past, the present and the future of population genomics in this species. Tomato is a model species in genetics, as well as in population genomics thanks to the important collection of genomic data that have been accumulating over years. By highlighting the importance of crop wild relative species for adaptability improvement of modern cultivars, this chapter describes the scientific context of this thesis work.

In the second chapter, we aimed to decipher the most likely domestication scenario for the three crop and wild population pairs. We performed a comparative analysis of several demographic models of increasing complexity to limit biases induced by making strong assumptions (Gaut et al. 2018). Comparing the crop and wild populations enabled us to evaluate the extent of biological changes due to domestication. This knowledge is crucial to improve future breeding efforts and bring valuable estimation of the impact of human selection on the crop effective population size and gene

flow with their wild relative (Zeder 2015). Inferring the demographic scenarios of these three species is an unprecedented opportunity to further characterize each domestication event duration, and therefore improve the inference of the demographic history that were hypothesized through indirect means (human and cultivation history of the areas, ancient written records). This information is not described in the literature.

In the third chapter, the hypothesis relies on a convergent modification of gene nucleotide diversity and gene expression levels during domestication between the three species. Comparing crop and wild relative accessions enabled to estimate gene expression differences and detect genomic selection footprints. Annotations of the targeted genes (selected and differentially expressed) identified the biological processes altered during domestication. The hypothesis relies on the orthologs shared within the trio of species and their modification. We hypothesize that mechanisms of regulation and adaptation that have been triggered by domestication of crop species are convergent. Therefore, for the three independent domestication process the expectation is to highlight parallel changes induced in crops compare to their wild relatives.

# MATERIALS AND METHODS

## a. Data available before the start of the PhD project

It has been feasible to start such a project for my PhD only thanks to the dataset already available. The tomato samples were part of the ARCAD project (project No 0900-001 supported by The Agropolis Fondation), and partly published in Sauvage et al. (2017). This project aimed to explore the effect of domestication on genome evolution in 13 crops including the tomato. RNAseq data were produced for 10 crop (*S. lycopersicum*) and 10 wild (*S. pimpinellifolium*) accessions. Following up these analyses, the SOLUTION project attributed to Christopher Sauvage (EU Marie Curie Career Integration grant: FP7-PEOPLE-2011-CIG grant agreement PCIG10-GA-2011-304164) aimed to produce a comparative analyses of domestication effects within the *Solanaceae* family where the preliminary idea was to sequence the transcriptome (RNAseq) similarly to the ARCAD project, of 24 accessions including crops, wild species and a supplement of several outgroup species. In eggplant, the RNAseq data set included 6 crop accessions (*S. melongena*), 6 semi-domesticated accessions (*S. melongena* group E and G), 9 wild accessions (*S. melongena* group E and F) and 2 outgroups accessions (*S. incanum*); all these species determination followed the taxonomy from (Lester and Hasan 1990). In pepper, the RNAseq data set was composed of 9 crop accessions (*C. annuum*), 7 presumably wild relative accessions (*C. annuum var. glab*) and 8 accessions from 5 outgroup species (*C. microcarpum, C. frutescens, C. chinense, C. chacoense, C. baccatum*). In tomato, with the availability of the 20 ARCAD accessions, the SOLUTION RNA sequences aimed to explore further close wild relative species including 8 accessions from the Hirsutum group (1 *S. hirsutum*, 4 *S. habrochaites* and 3 *S. pennellii*), 7 accessions from the Peruvianum group (2 *S. peruvianum*, 2 *S. corneliomulleri*, 2 *S. huaylasense* and 1 *S. chilense*), 6 accessions from the Arcanum group (1 *S. arcanum*, 2 *S. chmielewskii* and 3 *S. neoricki*) and 3 accessions (*S. chesmanii*) from the Esculentum group, common to *S. lycopersicum* and *S. pimpinellifolium*. Outgroup species were used to improve polarization rate of SNP (ancestral *vs* derived state) to further unfold AFS. For outgroups species in eggplant and pepper, accessions were selected within the GR of the CRB-leg seed bank located at the UR1052 GAFL research unit. The choice was made according to the known divergence and taxonomic position inferred from Carrizo García et al. (2016) for pepper and from Aubriot et al. (2016) for eggplant. I extracted available RNAseq data of several wild species of tomato (Appendix 1) but mostly of 2 outgroup species (2 accessions from *S. lycopersicoides* and 1 accession from *S. sitiens*), from a published analysis from Pease et al. (2016). However, concerning the wild relative species, we chose

not too genetically distant species from the focus crop species to avoid decrease in mapping accuracy (details are developed in the table 3).

### b. Choice of plant accessions

The accessions studied included wild, domesticated and outgroup species part of the GAFL genetic resources and for each species a selection was provided according to known phylogenetic relationship and molecular data (mainly from SSRs genotyping; tomato (Roselius et al. 2005; Tam et al. 2005, 2007; Labate et al. 2007; Ranc et al. 2008), pepper (Paran and Van Der Knaap 2007), eggplant (Frary et al. 2000; Nunome et al. 2001)), to cover the widest range of nucleotide diversity. Therefore, the accessions sequenced afterwards were selected for the genetic diversity within each population of wild or domesticated plants.

The total material is composed of 92 samples among the three species (detailed description Appendix 1). From these data set we selected accessions to perform the analyses. In the tomato data sets I studied only 3 species (9 accessions of *S. lycopersicum*, 9 individuals of *S. pimpinellifolium* the close wild relative and 12 accessions of a further apart species *S. peruvianum*) and 3 outgroup individuals (*S. lycopersicoides* and *S. sitiens*). In pepper the data set was composed of 6 accessions of *C. annuum* (and reclassed C. *annuum var. glabriusculum*) and 4 accessions of the close wild relative *C. annuum var. glabriusculum*, 4 accessions of *C. frutescens* and *C. chinense*, and 4 accessions outgroup of *C. microcarpum*, *C. baccatum* and *C. chacoense*. And finally, the eggplant data set included 7 accessions of *S. melongena* (including an accession of *S. insanum* close from the *S. melongena*), 11 accessions of *S. insanum* and 2 accessions outgroup of *S. campylacanthum.*

In the figure 12, the principal component analyses graphically represent the genetic distances between each accession of the three species. All outgroups are present in the figure 14*a* and they all separated clearly from the crop and wild relative species. Eggplant accessions considered as semi-wild had to be reclassified into a new species as *S. insanum* was not yet considered as a species when the accessions were sampled. To proceed, I followed the advices of the eggplant taxonomy expert Dr. Xavier Aubriot(Aubriot et al. 2018). The eggplant shows a continuum of genetic changes from the crop to the wild accessions. To proceed to the demographic inferences, it was necessary to have two clear genetically distinct groups without a strong structure, and it explains the differences in accession choices for the chapter 2 and the chapter 3 (See table 3 and figure 12*b* & 12*c*).

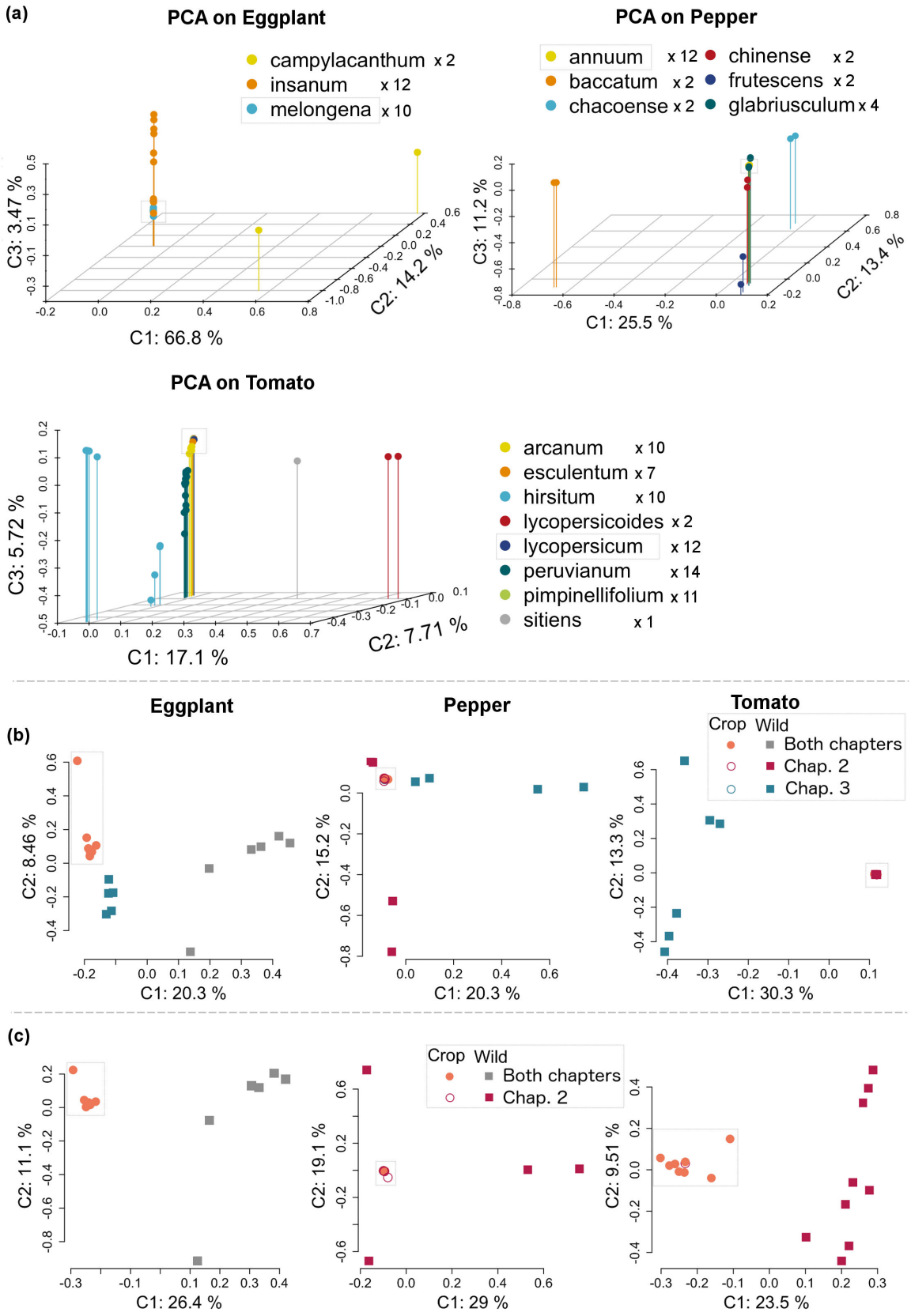



**Figure 12**. Graphical representations of the principal component analyses of the three species accessions, to facilitate the reading, crop species are in-boxed in each PCAs. (a) PCAs of the total accessions available, colored according to their species for eggplant and pepper, and to their groups for tomato. (b) PCAs of the accessions used for the thesis work analyses, circles: crop accessions, square: wild accessions, colors referring to the chapter using these accessions. (c) PCAs of the accessions used for the chapter 2.

The wild pepper accessions were chosen in the genetic resources available to explore the wider genetic diversity possible, this might explain the strong structure within the 4 close wild relatives *C. annuum var. glabriusculum* which appeared after analyses to be part of different, not yet described as separate, species (and as mentioned in the *§IV.a.* two of the accessions were reclassified to *C. annuum*). The second chapter aimed to better understand the domestication demography and required the closest relative species to have a reliable resolution on the domestication process. But in the third chapter, we chose to perform the transcriptomic and the nucleotide diversity analyses at the gene scale and to increase the statistical power we decided to include the two species *C. frutescens* and *C. chinense*. These further apart relative species description of differentially expressed genes. The tomato accessions number was higher as we used previous work on tomato to complete the analyses (Pease et al. 2016; Sauvage et al. 2017), we chose to work on the closest wild relative *S. pimpinellifolium* in the chapter 2. Though, after the publication of Sauvage et al. (2017) on the transcriptomic rewiring between *S. lycopersicum* and *S. pimpinellifolium* and with a purpose to avoid redundancy and complete available knowledge on transcriptomic changes due to domestication, the chapter 3 focuses on the group peruvianum, details in the table 3 and representation in the figure 11*a*.

**Table 3**. Details of the accessions chosen for both the chapter 2 and 3.

| | | Species | Number used in Chapter2 | Number used in Chapter3 |
|---|---|---|---|---|
| Eggplant | Crop | *S. melongena* | 7 | 7 |
| | Wild | *S. insanum* | 6 | 11 |
| | Outgroup | *S. campylacanthum* | 2 | - |
| Pepper | Crop | *C. annuum* | 10 | 8 |
| | Wild | *C. annuum var. glabriusculum* | 4 | - |
| | | *C. chinense* | - | 2 |
| | | *C. frutescens* | - | 2 |
| | Outgroup | *C. baccatum* | 2 | - |
| | | *C. chacoense* | 2 | - |
| Tomato | Crop | *S. lycopersicum* | 9 | 8 |
| | Wild | *S. pimpinellifolium* | 9 | - |
| | | *S. peruvianum* | - | 2 |
| | | *S. huaylasense* | - | 2 |
| | | *S. corneliomulleri* | - | 2 |
| | Outgroup | *S. sitiens* | 1 | - |
| | | *S. lycopersicoides* | 2 | - |

### c. Preparation of the biological material

The plants were grown in a greenhouse in spring 2012 at INRA with required environmental condition (watering, sun day light and temperature regulation) for each species to avoid biases in gene expression levels. For example, young leaves tissues were sampled at the same hour of the day, across accessions plant tissues were sampled (3 replicates per accession), flash frozen in liquid nitrogen prior to the production of the RNAseq libraries for each sample as follows: sampled tissues were pooled according to a 15, 20 and 65% proportion of flower, fruit and fresh leaves , respectively to represent equal amount in µg of RNA, to get the best representation of the gene expression levels in every plant organ and be consistent with the biological material produced in the framework of the Arcad project. Fruit samples were harvested at the ripe stage (40 days post-anthesis) of each species. Then RNA was quantified and qualified using a bioanalyser. RNAseq libraries were prepared and individually tagged using a 6 bp tag at INRA SupAgro (Montpellier) using the TrueSeq kit and sequenced using the HiSEQ2500 protocol (150bp orientated paired-end reads) from the Genotoul Platform (INRA, Toulouse).

### d. Alignment of the RNAseq data set

The analyses were based on the RNAseq data of all the accessions listed above of the three *Solanaceae* species. The strong advantage of RNAseq data is that we could both analyze the expression of the genes and their genetic diversity (on the coding regions only). In order to process these data, we built a bioinformatic workflow (*cf* detailed bioinformatic workflow for software and parameters p 53-56) that is composed of the classical major steps including quality control, mapping and the SNP detection that is the center of this thesis work because the inferences, the transcriptomic and the diversity analyses depend on the mapping quality and the variant calling.

The mapping was performed against the version ITAG3.2 (The Tomato Genome Consortium 2012) of the tomato transcriptome, the v2.0 (Qin et al. 2014) of the pepper transcriptome and was initially done on the draft genome of the eggplant (Hirakawa et al. 2014). At first, with the draft genome, the high number of contigs limited the approach. By later accessing the eggplant reference genome (not yet published) (, the mapping accuracy increased and allowed the use of the eggplant data set for the comparative analyses (e.g. ortholog analysis between the three reference genomes).

### e. Demographic inference modeling

In a first part, we used outgroup species to polarize the polymorphisms detected between crop and wild populations (figure 13). Using a summary of the whole population genetic diversity (jSFS), we tested over ~40 inference models, that were run 50 times independently to offer consistency in the results. Hypotheses included were: strict isolation or isolation with migration that would experience 1, 2 or 3 demographic events, with or without bottleneck, constant or increasing/decreasing effective population size (*Ne*) at each step. And we completed the analyses by adding the possibility to have heterogeneous variation of *Ne* across the genome (selective sweep), figure 6*a*, or, heterogeneous variation of migration across the genome (selection against migrant), figure 6*b*, or both. Some models had poor score (low maximum likelihood on all runs and, therefore, were discarded from the final analyses. After selection of the 10 models that would cover the widest range of scenario possible (e.g. effective size expansion, bottleneck, unique or multiple demographic events etc.), I ran analyses presented in the chapter 3. This comparative method allowed the unbiased choice of the most probable scenario (on the basis of the maximum likelihood criteria) of demographic history for the crop and wild populations of the three species. To ascertain the choice of best demographic model, following a recent example study (Fulgione et al. 2018), we selected the second best scenario and compared the parameter estimations between each of these scenarios. As expected, even with different demographic scenarios, the parameters converged towards consistent estimations.

The whole genomic diversity is impacted by domestication especially due to changes in:

- recombination rate (reducing the linked selection) which we chose to ignore by removing LD sites,
- mating system, with an increase of inbreeding to conserve fixed traits within cultivars, which is common to all Solanaceae domesticated species,
- and demography that impacts the effective population size.

Understanding the divergence between crop and wild populations and the course of the domestication process via the characterization of demographic events in crop species is essential to better understand this evolutionary process. This part is developed in the Chapter 2. The bioinformatic workflow is detailed in the GitHub repository located at https://github.com/starnoux/arnoux_et_al_2019.

### f. Gene expression analyses

Based on the results reported in Sauvage et al. (2017), we extended the approach and tested for the parallelism/convergence in the imprinting of domestication on the landscape of gene expression levels. Domestication can be studied through differences in gene expression between crop and wild populations. Though many levels of regulation affect transcription and this not only on a sequence variation manner, we coupled these analyses to ortholog analyses and common population genetic estimators. The aim was to detect convergent or divergent selective and transcriptional footprint of domestication. We performed a transcriptomic comparative study that revealed genes differently expressed and their correlation with the footprints of selection on the genetic diversity loss and gain across the expressed genes. This part is developed in the Chapter 3 and the bioinformatic workflow is detailed in the GitHub repository {https://github.com/starnoux/arnoux_et_al_2018}.

### g. Complementary details on the bioinformatic workflow (p 53-56)

#### i. Common bioinformatic workflow to both chapter analyses

- Controlling for the quality of the raw sequencing data and removal of the lowest quality reads
- Mapping the RNAseq reads to the reference genome, and insuring no bias is affecting the mapping accuracy across individuals (discrepancy due to genetic divergence with the reference genome) and along the genome (gene paralogs).
- Calling for SNPs, at this step, is crucial to make sure that polymorphisms detected are real instead of an artefact due to paralog genes (i.e. homologous genes that separated because of gene duplication events). Basically, if two genes are similar, the reads might map to each other and the few changes would be considered as polymorphisms when they are only reflecting the presence of two paralogs. To ensure the quality of the SNPs, we filtered the potential paralog sites with the method implemented in Reads2SNP (Nabholz et al. 2014).

#### ii. Comparing inference modeling

- The SNPs were then filtered (LD pruning) to perform demographic inferences (with ðaði), as there is an assumption of independency of the SNPs. I performed this filtering to insure we were fulfilling the requirement for the demographic inferences, and to avoid redundant information brought by linked SNPs.

- The demographic inferences were then performed on site/allele frequency spectrum, used to describe the amount of genetic variation across the expressed genes in each species. It is a statistical summary of the polymorphisms of a population (See figure 13*a*). By performing a joint site frequency spectrum allele (jSFS) between the crop and the wild population, we could detect the shared polymorphisms and the frequency of each SNPs within one or both populations (depicted as purple dots in the figure 13*b*).
- The figure 13*c*, details the method implemented in ðaði software to estimate the different demographic parameters such as the genetic drift in the crop and the wild populations or the migration (asymmetric gene flow) from the study of the jSFS. The inferences aim to determine if the given model fits better the observed data.

### iii. Transcriptomic, ortholog, gene ontology and nucleotide diversity analyses.

- The summary statistics (π and Tajima's D) for nucleotide diversity and demography were produced with DNAsp. Briefly, the nucleotide diversity (π) is a relative measure of the degree of polymorphism within a population that can be used to detect balancing or directional selection and hard sweeps (Hohenlohe et al. 2011). Tajima's D is the difference between $\theta_\pi$ and $\theta_w$ (the observed diversity against the expected nucleotide diversity) and estimates both evidences of selection (equilibrium, selective sweep or balancing selection) and the demography of a population (neutrally evolving population, population expansion after a recent bottleneck or population contraction). In the last chapter we used a complex of species for the wild tomato and pepper, therefore the Tajima's D could not be used as the two species may have experienced different demographic events that would impact the estimator. Though, while using π, at the gene level, we could scan for chromosomic regions under selection. Strongly selected genes are expected to have low π, thus, comparing the crop and wild, the changes in nucleotide diversity reveal genes experiencing selective pressures of selection during domestication.
- The Differentially Expressed genes (DEGs) were detected on normalized gene expression within population. Both crop and wild accessions were clustered and the mean expression of each gene was compared between populations to reveal under- or over-expressed in the crop population.
- To foster the biological interpretation of the DEG, and to avoid heterogeneity in the genes annotations across the 3 species, we used the protein family database Pfam to annotate the reference proteomes (the translated coding sequences (CDS) of the reference genome) with the UniProtKB (database of all coding protein identified in all species). These annotations allowed the

detection of the gene family, and, the processes and pathways they are involved in but we focused only on the biological process for our studies.

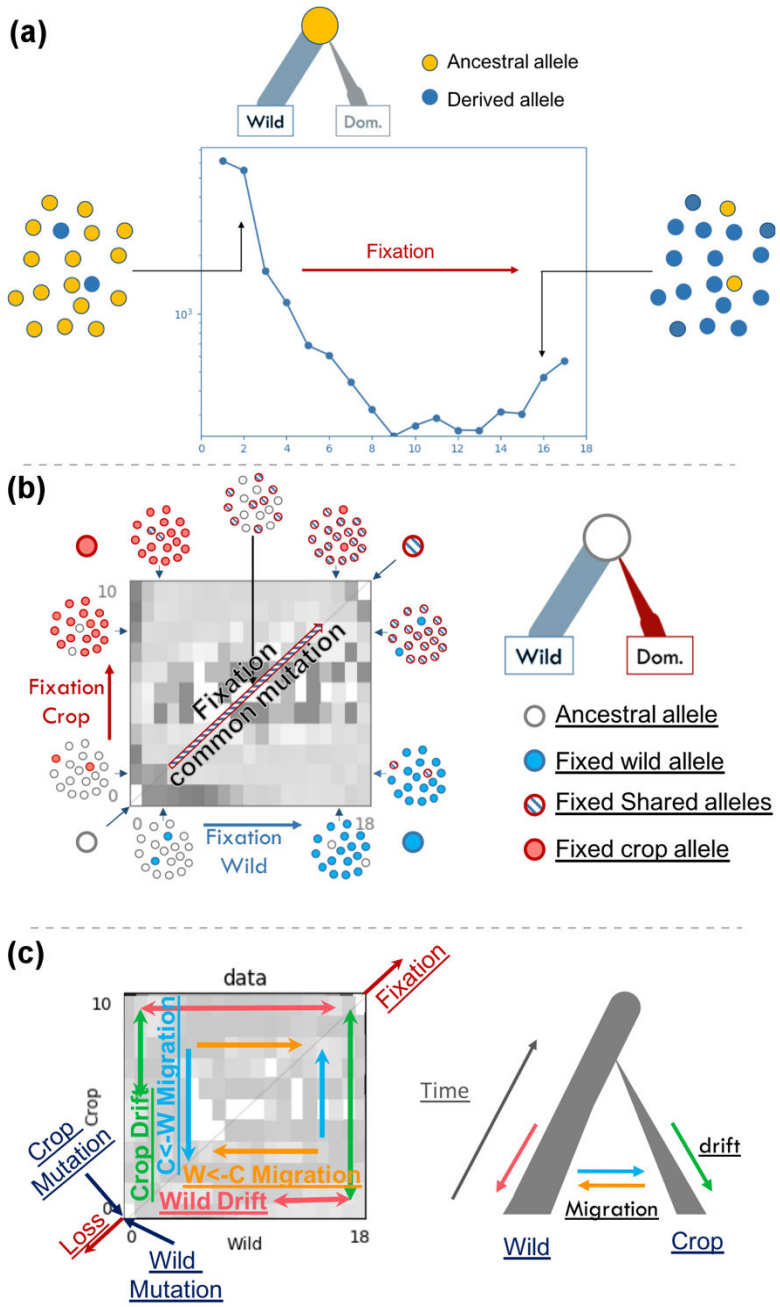



**Figure 13**. Demographic inferences from a joint site frequency spectrum

(*a*) Site frequency spectrum of a population of 9 individuals (diploid). The cluster of dots represent the frequency in ancestral and derived alleles at the population level, for one site.

(*b*) These joint site frequency spectra are based on a heatmap representing the shared and species-specific, derived or ancestral alleles.

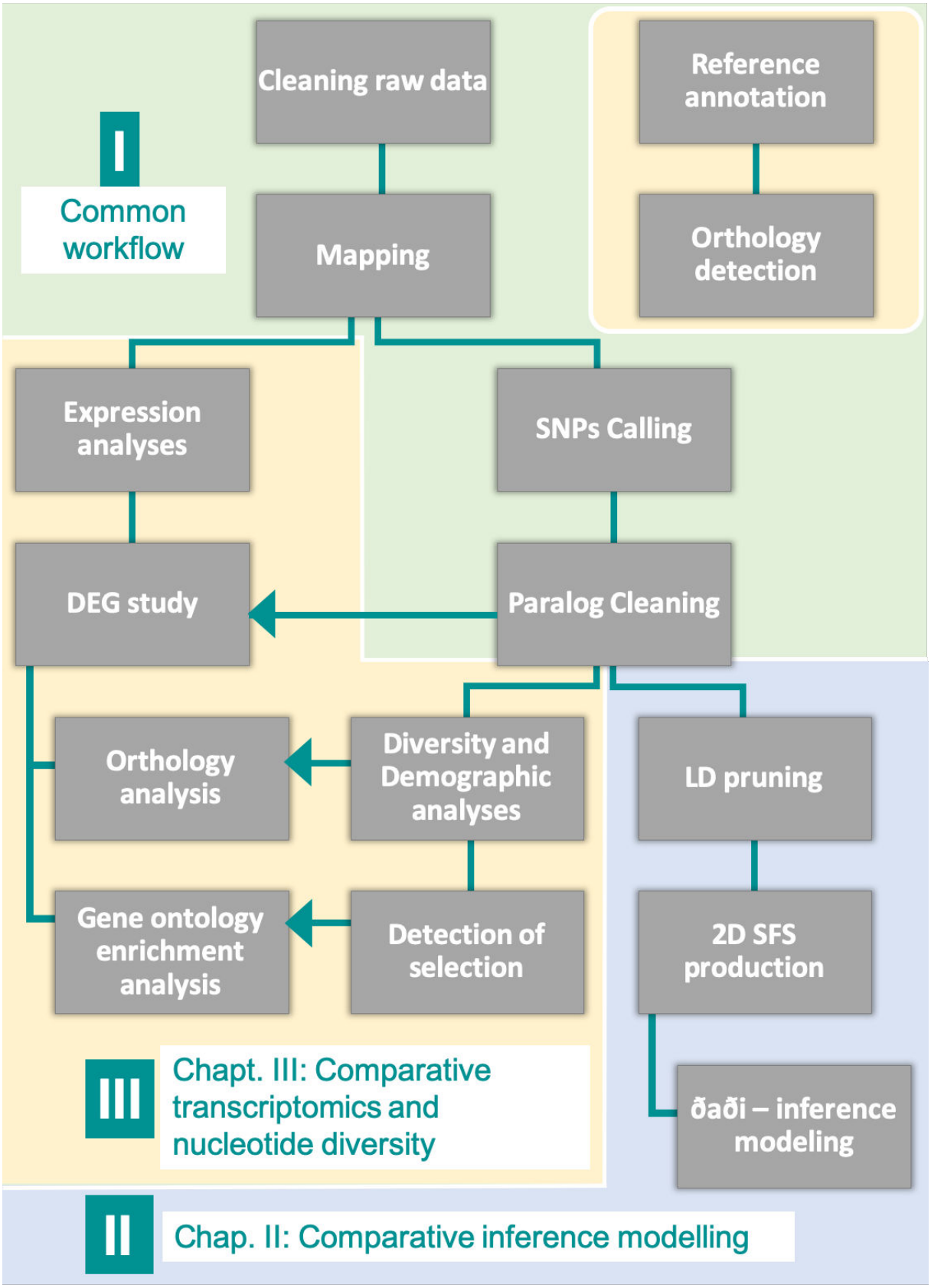
(c) The joint site frequency spectrum and the significance of each area translated to a demographic tree on the right side.

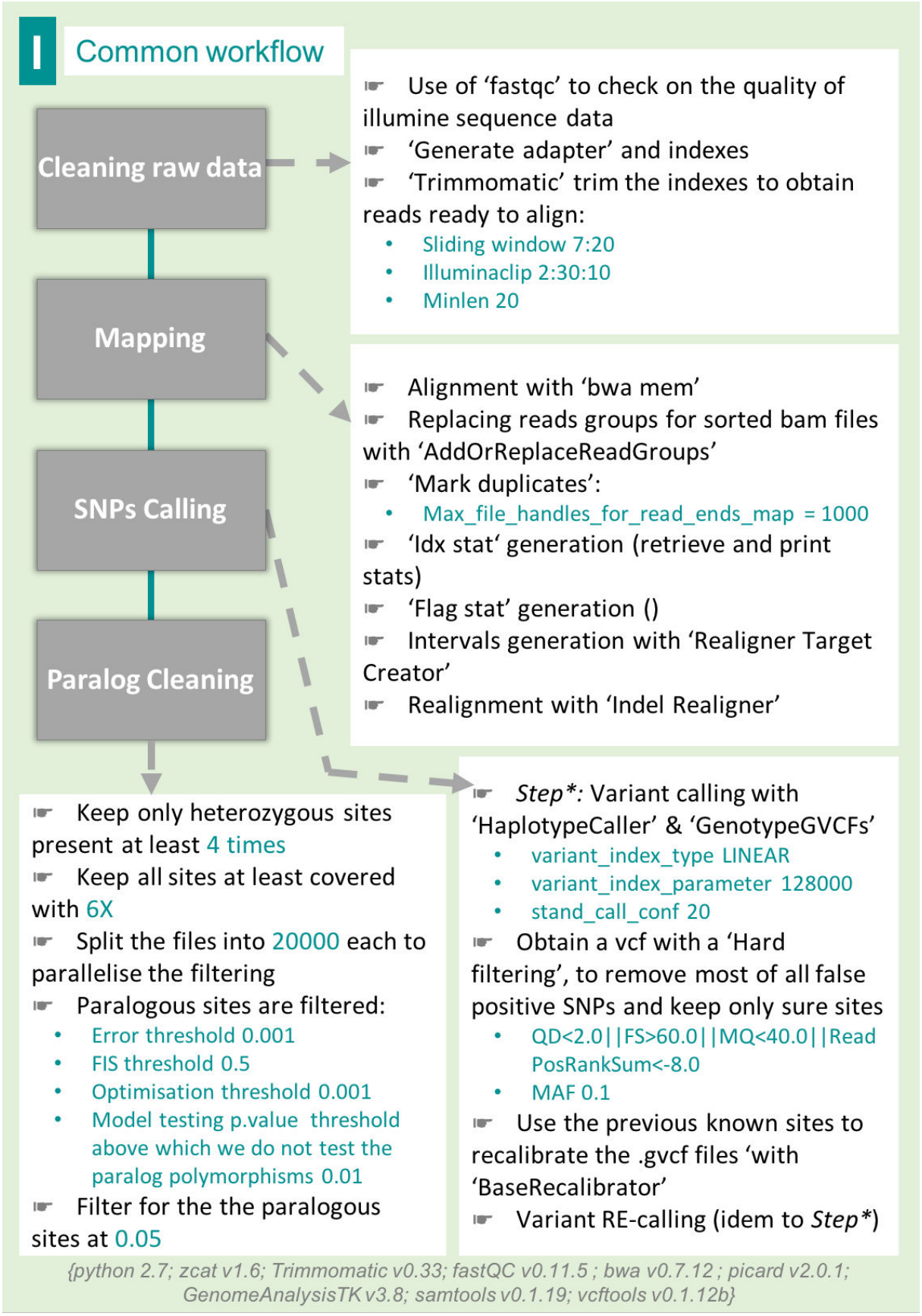
*Inspired by: Gutenkunst et al. 2009*

- The gene ontology analyses performed on the annotated genes aims to define the gene function over-represented within a set of given genes (e.g. genes under selection, gene differentially expressed). The GO analyses give a representation of the processes and pathways modified during domestication, when comparing the gene ontology of the crop selected genes within each species. These results highlight parallel and convergent domestication footprints, but without dissociating the two phenomena.

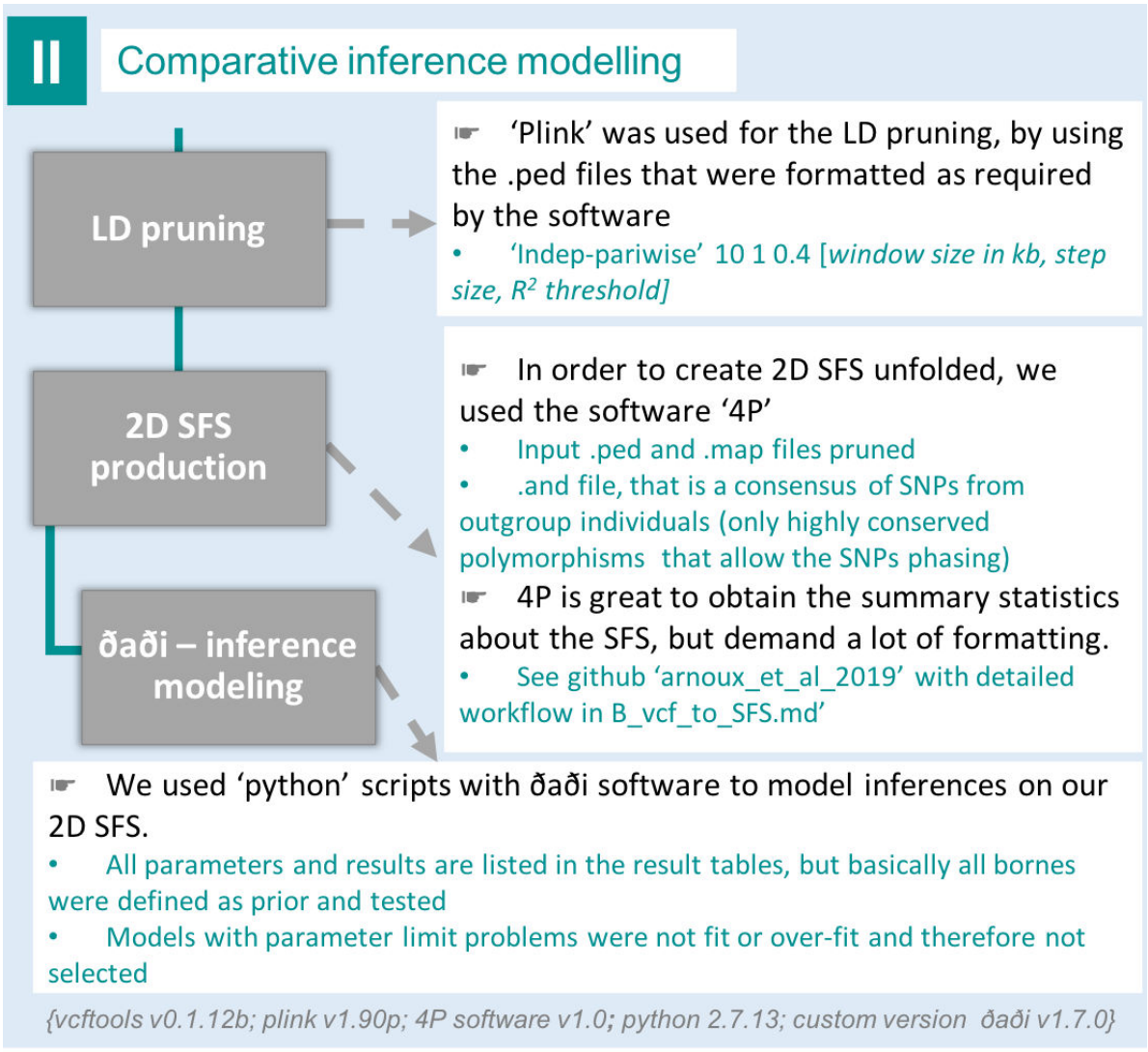
- The ortholog analysis was performed to ascertain how many and which genes were similar between the three species. It compares the protein coding genes pairwise and establishes if they are orthologs for two or the three species of interest. In this case, the three species are from a same plant family, therefore quite close genome-wise which facilitates the analyses. Finding similarly selected or differentially expressed genes reveals a convergence of domestication footprints and dissociates it from a parallel one.

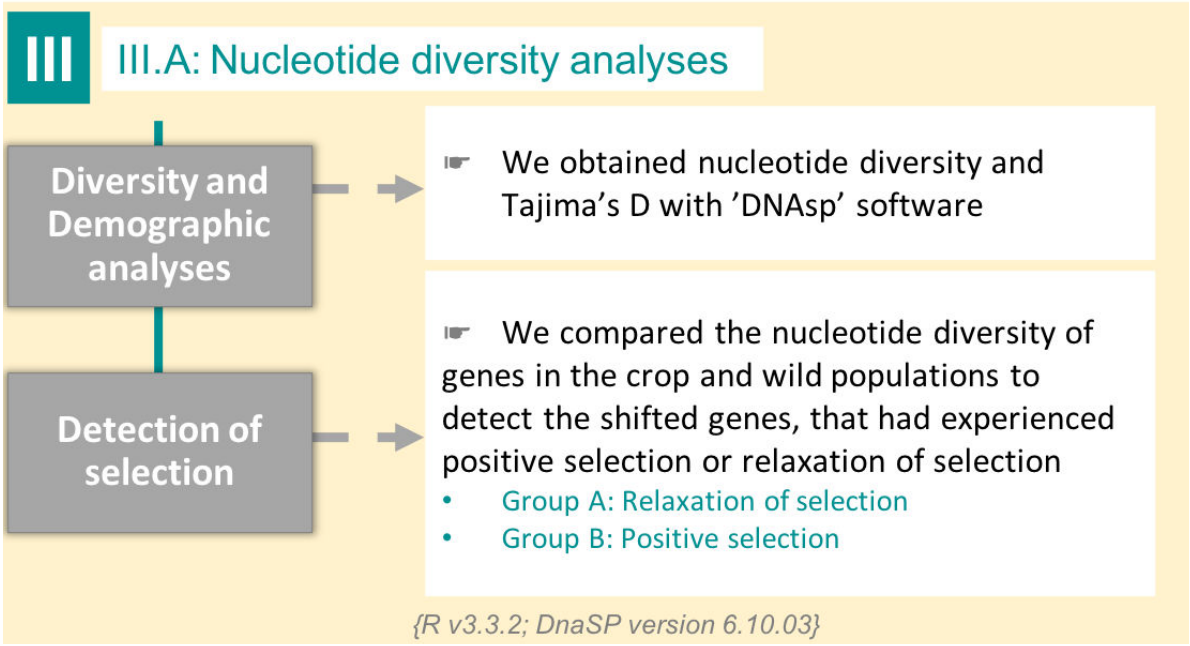
## DETAILED BIOINFORMATIC WORKFLOW

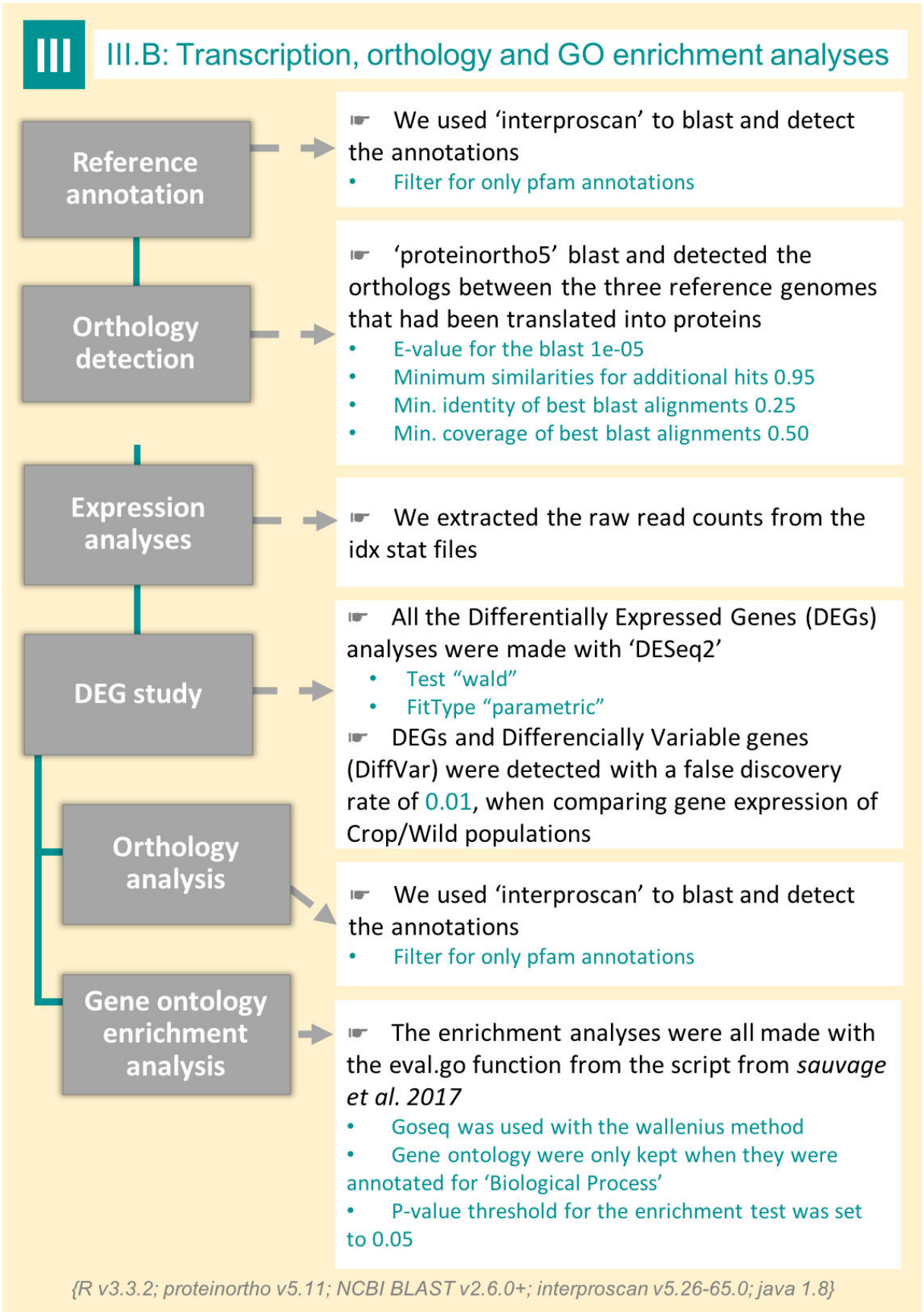





*{python 2.7; zcat v1.6; Trimmomatic v0.33; fastQC v0.11.5 ; bwa v0.7.12 ; picard v2.0.1; GenomeAnalysisTK v3.8; samtools v0.1.19; vcftools v0.1.12b}*


II
Comparative inference modelling
LD pruning
2D SFS production
ðaði – inference modeling
☛ 'Plink' was used for the LD pruning, by using the .ped files that were formatted as required by the software
• 'Indep-pariwise' 10 1 0.4 [window size in kb, step size, R² threshold]
☛ In order to create 2D SFS unfolded, we used the software '4P'
• Input .ped and .map files pruned
• .and file, that is a consensus of SNPs from outgroup individuals (only highly conserved polymorphisms that allow the SNPs phasing)
☛ 4P is great to obtain the summary statistics about the SFS, but demand a lot of formatting.
• See github 'arnoux_et_al_2019' with detailed workflow in B_vcf_to_SFS.md'
☛ We used 'python' scripts with ðaði software to model inferences on our 2D SFS.
• All parameters and results are listed in the result tables, but basically all bornes were defined as prior and tested
• Models with parameter limit problems were not fit or over-fit and therefore not selected
{vcftools v0.1.12b; plink v1.90p; 4P software v1.0; python 2.7.13; custom version ðaði v1.7.0}


III
III.A: Nucleotide diversity analyses
Diversity and Demographic analyses
Detection of selection
☛ We obtained nucleotide diversity and Tajima's D with 'DNAsp' software
☛ We compared the nucleotide diversity of genes in the crop and wild populations to detect the shifted genes, that had experienced positive selection or relaxation of selection
• Group A: Relaxation of selection
• Group B: Positive selection
{R v3.3.2; DnaSP version 6.10.03}




*{R v3.3.2; proteinortho v5.11; NCBI BLAST v2.6.0+; interproscan v5.26-65.0; java 1.8}*

# CHAPTER 1

## Progress and prospects of population genomics in major crop plants - Tomato population genomics

In the first chapter, the state of the art of research on tomato as model species gives an overview on the past, the present and the future of population genomics in this species. Tomato is a model species in genetics, as well as in population genomics thanks to the important collection of genomic data that have been accumulating over years. By highlighting the importance of crop wild relative species for adaptability improvement of modern cultivars, this chapter describes the scientific context of this thesis work.

**Book:** Population Genomics: Concepts, Approaches and Applications.

**Section:** Progress and prospects of population genomics in major crop plants

# **Chapter:** Tomato population genomics

Manuscript under revision

Stéphanie Arnoux[1], Mathilde Causse[1], Christopher Sauvage[1]*

[1], INRA UR1052 GAFL, Centre de Recherche INRA PACA, Domaine Saint Maurice, 67 Allée des Chênes, CS60094, 84140 Avignon Cedex 9, France

* Corresponding author

Om P. Rajora (ed.), Population Genomics: Concepts, Approaches and Applications,

Population Genomics [Om P. Rajora (Editor-in-Chief)],

© Springer International Publishing AG, part of Springer Nature 2018

# Chapter 1

## *Progress and prospects of population genomics in major crop plants - Tomato population genomics* *- 57 -*

**Abstract** **- 61 -**

**Introduction** **- 63 -**

**Part I: How tomato became the model plant for vegetables** **- 64 -**
1. Tomato history, from past to modern era ..... - 64 -
2. Towards the reference genome of tomato and databases ..... - 65 -
3. Genome and transcriptome sequencing of crop wild relative species ..... - 66 -

**Part II: Tomato as a model for Molecular Evolution** **- 67 -**
1. Original organization of the Tomato clade ..... - 67 -
2. Modern phylogeny and taxonomy of the Tomato clade ..... - 68 -
3. Ecological Genomics of the tomato crop and its wild relatives ..... - 70 -
4. Genomic footprints of domestication and modern breeding stages ..... - 76 -

**Part III: Population genomics to sustain modern breeding** **- 79 -**
1. Introgressions from crop wild relative species improved the crop tomato ..... - 80 -
2. Dissecting the genetic architecture of agronomical traits ..... - 81 -
3. Molecular bases of trait diversification ..... - 82 -
4. Breeding shaped the genetic structure of modern cultivars ..... - 82 -
5. Genome-wide association approach extended the knowledge of the genetic architecture of agronomical traits - 84 -

**Part IV: Prospects for future research** **- 87 -**
1. Towards a pan-genome in tomato ..... - 87 -
2. Modelling of demographic history and ecological niche ..... - 87 -
3. Adapting the tomato crop to climate change using genomic approaches ..... - 89 -
4. Implementing genome-wide based Genomic Selection ..... - 89 -

Further Major Readings We Recommend: ..... - 92 -

## Abstract

Tomato is an acknowledged model species for research in genetics and genomics, on fruit development and disease resistances, but it also deserves to be a model for population genomics thanks to the large genetic and genomic resources available. Tomato improvement largely depends on introgressions of beneficial alleles from wild relative species.

Since the first release of a high-quality genome sequence of the tomato crop in 2012, the genomes of several hundreds of cultivated accessions and a few wild relatives have been re-sequenced, allowing the discovery of millions SNP. Their study confirmed the new phylogenetic organisation and the monophyletic origin of the *Solanum* genera section *Lycopersicum*, composed of 13 species. Recent ecological genomics approaches, notably using RNAseq approach, provided new results on speciation and interspecific reproductive barriers. The molecular mechanisms of adaptation to abiotic stress in crop and wild tomatoes were also analysed and their role underlined as factors of speciation and diversification. The diversity of ecological conditions of the wild relative species allowed the study of evolutionary and molecular mechanisms of adaptation to abiotic stress in crop and wild tomatoes.

Using genomic studies, the two steps of tomato domestication and the intensity of bottlenecks due to domestication and further modern breeding were clarified. Selection footprints and large genomic regions introgressed from the wild relative species were identified. At the transcriptome level, it was also shown that domestication and modern breeding rewired genome expression, notably for stress related genes.

Finally, the availability of genome sequences and SNP markers allowed studying large collections of varieties, developing GWAS and advancing our knowledge about the genome structure (linkage disequilibrium decay, distribution of recombination), but also mapping genes and QTL involved in many traits and using the information for breeding new varieties.

## Introduction

Tomato is the first vegetable grown over the world. It accounts for more than 15% of the world vegetable production (over 177 million metric tons in 2016; Food and Agriculture Organisation [faostat 2016]). Half of the world production is produced in four countries (China 56 MT, India 18 MT, USA 13 MT and Turkey 12 MT). Tomato is grown for two main usages: processing and fresh market. It is a rich source of micronutrients in human diet. The major goals of tomato breeding (high productivity, tolerance to biotic and abiotic stresses and high sensory and health value of the fruit) require a good comprehension and management of tomato genetic resources and diversity. Tomato is also an acknowledged model species for research in genetics, on fruit development and disease resistances. It has a short life cycle, is easy to cross and self-pollinate in its crop form, it has a medium size genome (approximately 900 Mb) and large genetic and genomic resources. Furthermore, the tomato scientific community has access to several databases gathering most of the important data.

Tomato and its 12 closely related species belong to the Solanum genus in the large Solanaceae family. All the species come from the Andean region of South America. Explorations of tomato centre of origin permitted major advances in the characterization of its genetic and phenotypic diversity. In parallel, *ex situ* conservation of genetic resources in large national collections ensured the conservation of landraces and wild species. Thus, the genetic potential of tomato's wild relatives for breeding purpose emerged. In parallel, the ecological and taxonomic diversity of tomato turned it into a model species for evolutionary studies. Since the mid-20th century, mastering controlled hybridization allowed crosses between wild and cultivated tomato to be performed. Modern genetics and breeding methods contributed to understand the genetic control of agronomical traits but also accentuated the progress and the development of thousands of new cultivars. It also underlined the value of crop wild relatives.

The advent of molecular biology in the 80's raised great hopes in terms of characterization of the genetic diversity present in both wild and cultivated compartments. Great expectations also emerged since the development of molecular techniques to pinpoint genomic regions involved in targeted traits. Dissection of the genetic control of complex traits, using ad hoc techniques from quantitative and population genetics, was possible, leading to the identification of key alleles involved in many traits, originating from several wild relatives. Today the tomato genome is fully sequenced and the genomes of many wild and cultivated accessions have been re-sequenced thanks to high-throughput sequencing (HTS) techniques. Large datasets describing the genome expression

(transcriptome, proteome and metabolome) are also available providing an overview of the (post-)transcriptional landscape. Quantitative trait locus (QTL) mapping techniques or genome wide association studies (GWAS) also facilitate the understanding of the genetic architecture of complex traits and germplasm management of both wild and cultivated tomatoes.

In this chapter we first describe the tomato history, its domestication and the diversity and phylogeny of its wild relative species. We then present the genomic resources available on the clade and how they have provided new insight on the evolution and diversity of tomato accessions. We then focus on the impact of domestication and breeding, before to show how crop wild relatives were used to introgress and identify important loci for the crop. Finally, some major prospects are proposed.

## Part I: How tomato became the model plant for vegetables

### 1. Tomato history, from past to modern era

Tomato (*Solanum lycopersicum* L.) and its 12 wild relative species are originated from the Andean region of South America (de Candolle 1886; Jenkins 1948; Rick and Fobes 1975; Spooner et al. 2005; Peralta et al. 2008; Zuriaga et al. 2009). Its common name 'tomato' originates from the Nahuatl (Aztec language) world 'tomatl'. The origin of domestication was debated over the years but recent studies untangled this mystery. Briefly, it was first domesticated from the wild species *S. pimpinellifolium* by ancestors of Inca population in Ecuadorian and Peruvian regions. The beginning of trade between populations from South- and Mesoamerica later introduced few individuals in the Mexican region leading to a strong bottleneck (Blanca et al. 2012). A second strong bottleneck occurred with the Spanish colonization of the American continent when Mesoamerican tomato seeds were brought to Europe. Tomato started to be consumed in Europe as food during the 17th and 18th century, and in 1869 Henry John Heinz founded the first company linked with tomato (Ray 1673; Labate et al. 2007).

Since then, the tomato spread worldwide and in the early 1920's a field of tomato improvement research appeared to obtain the first disease tolerant cultivars from hybridization with wild progenitors. The first resistant cultivars to *Cladosporium fulvum* and *Fusarium oxysporum* appeared in the 1930's and 1940's with the discovery of resistance genes in the closely related wild tomato species (Langford 1937; Stevens and Rick 1986). From then on, tomato improvement has largely been dependent on introgressions of beneficial alleles from wild germplasm (Atherton and

Harris 1986) which increased the interest for the knowledge and conservation of crop wild relative species. After the pioneer Nikolai Vavilov (Kurlovich et al. 2000), the main protagonist in the development of a crop and wild seed bank was Charles M. Rick who dedicated his life in field trips in South America and established the Tomato Genetics Resource Center (Rick 1990, https://tgrc.ucdavis.edu/). These efforts, to increase crop and wild tomato sampling and making it available to the scientific community, are part of the reasons that brought the tomato up to a 'model species' position. The other reason to deepen the research in tomato is its economic importance as one of the leading vegetable crops worldwide. As a reference in the past 25 years (1984-2014), the global yield of tomato increased from 83 to 170 million tons and the area harvested increased from 3 to 5 million hectares (United Nations Food and Agriculture Organization (FAO) statistics; Food and Agricultural Organization of the United Nations [faostat 2016]). The scientific and agricultural community considerably improved the tomato varieties and growth conditions in the last 50 years, notably for its yield, stress tolerance, fruit properties and pathogen resistances (Bauchet and Causse 2012).

## 2. Towards the reference genome of tomato and databases

In the early 2000's, the Tomato Genome Consortium was set up. It was an international consortium of scientists from 14 countries that gathered their funds to sequence the first tomato genome *(Solanum lycopersicum)* (among other Solanaceae species) and provide a resource publicly available. Following the first objective to sequence the 220 Mb of tomato euchromatin, predicted to contain the majority of genes (Mueller et al. 2005a), the next generation sequencing methodologies offered the opportunity to produce a mostly complete high-quality reference genome, that finally covered 742 Mb (i.e. 83% of the 900 Mb genome, Sato et al. 2012). This work is part of a larger initiative called the "International Solanaceae Genome Project (SOL): Systems Approach to Diversity and Adaptation". This community aims to help understanding the genetic basis of plant diversity by offering a big clade-oriented database within the Sol Genomic Network website (SGN, http://solgenomics.net/) that collects and stores all the Solanaceae and related species genome sequences, phenotypic and genomic data available (Mueller 2005; Menda et al. 2008; Bombarely et al. 2011). This database, available to all researchers, also implemented supplementary tools such as solQTL or SGN VIGS (Virus-Induced Gene Silencing) (Tecle et al. 2010; Fernandez-Pozo et al. 2015). In the same intent to create a tomato expression database, the Tomato Expression Atlas (Fernandez-Pozo et al. 2015, 2017), the Tomato EFP Browser, TomPLEX (Winter et al. 2007) and TomExpress

(Zouine et al. 2017) are now allowing browsing the transcriptional landscape of each annotated gene from different plant tissues, genotypes and conditions and displaying the results with graphical outputs.

Following the release of the first reference genome sequence of cultivated tomato (cultivar Heinz1706), the genome sequence and its annotation have been regularly updated from an initial version to the current one, the third (SL3.0) which integrated new whole genome shotgun, full-length BAC and optical sequencing and reduced the number of contig gaps. This reflects the efforts to offer a high-quality genome to reach the gold standard that is available since many years now in the model plant *Arabidopsis thaliana*.

### 3. Genome and transcriptome sequencing of crop wild relative species

Crop wild relative species are particularly useful in population genomics notably to polarise SNP markers (determine the derived/ancestral state to unfold site frequency spectrum), track introgression events for adaptive traits or help phylogeny to be rooted to understand the evolution of traits along (Farris 1982). The advent of the second and third sequencing generation technologies (i.e. Hiseq Illumina, long reads technologies, respectively) allowed reaching these objectives by first providing the complete genome sequences of several wild relative species of the cultivated tomato. Indeed, these technologies are more adapted for the outcrossing crop wild relative species to manage properly the higher level of heterozygosity compared to the crop tomato (due to their self-incompatibility). Among these crop wild relative species, the first genome of *Solanum pennellii* was sequenced using Illumina technology (LA0716 accession, Bolger et al. 2014) and was updated using a *de novo* assembly based on the nanopore technology (LYC 1722 accession, Schmidt et al. 2017). The main objective was notably to foster our knowledge of traits related to stress tolerance and the evolutionary role of transposable elements on these traits, as *S. pennellii,* endemic to Andean regions in South America has evolved to thrive in arid habitats. The genome completeness obtained from the *de novo* approach, compared to the previous version, illustrated the gain obtained from the use of long reads sequencing (i.e. Oxford Nanopore). Schmidt et al. were able to achieve assemblies for which the N50 contig length was 2.45 Mb (i.e. half of the assembly was in contigs of 2.45 Mb or longer) and the complete genome sequence was assembled in only 899 contigs. While being error prone, the estimated error rate, when using polishing software was similar to the Illumina technology, down to 0.025%. A complete reference genome was also released for the wild relative species *Solanum lycopersicoides*. Using the PacBio sequencing, with a coverage of 90x, the N50 and

genome coverage were estimated to 139kb and 89.7% (see https://solgenomics.net/organism/Solanum_lycopersicoides/genome). It should be noted that additional genomes are available for *S. pimpinellifolium* (LA1589 accession) and *S. galapagense* (LA0436 accession) but the assembled sequences are highly fragmented, limiting their use (http://ftp.solgenomics.net/genomes/Solanum_galapagense/).

In addition, re-sequencing efforts have been conducted to complement large-scale genomic panel studies for tomato mainly dedicated to population structure and GWAS (Aflitos et al. 2014; Lin et al. 2014) or investigate species barriers (Labate et al. 2014), providing data in crop wild relatives species. Besides complete and re-sequenced genomes, transcriptomic data produced at the genome-wide scale from crop wild relatives have been released in the last years. This approach is relatively powerful to reduce the complexity of the analysis by reducing the genome representation and cope with the higher level of polymorphism in these species. We can briefly mention RNAseq data from Pease et al. (2016) that produced reads across four clades (Esculentum, Arcanum, Peruvianum and Hirsutum), from Sauvage et al. (2017) in *S. pimpinellifolium* and from Florez-Rueda et al. (2016) and Beddows et al. (2017) in *S. peruvianum* and *S. chilense*. The main scientific results obtained from these data are detailed in the next sections of this chapter.

## Part II: Tomato as a model for Molecular Evolution

### 1. Original organization of the Tomato clade

The first botanist to consider domesticated tomato was Tournefort (1694), who recognized its close relationship with the genus Solanum but named the tomato genus *Lycopersicon* ("Wolf peach" in Greek). For a better nomenclature, Linnaeus (1753) intended to use consistently Latin binomials. He located the tomato in the *Solanum* genus and named the domesticated tomato *S. lycopersicum* and its wild relative *S. peruvianum*. The Gardener's and Botanist's Dictionary (Miller and Miller 1768) started using the Linnaeus' binomial system but kept the *Lycopersicon* genus and it is only in the 1807's edition that the tomato joined the *Solanum* genus. After these feeble taxonomic beginnings most of the taxonomists and gardeners kept the *Lycopersicon esculentum* name until the 1980's when the first phylogenetic studies started confirming the *Solanum* affiliations (Rick and Tanksley 1981; Spooner et al. 1993). The *Linnaeus* nomenclature has gained wide acceptance but *Lycopersicon* might remain present in the common language. The first phylogenetic studies brought a new growing interest in deciphering the crop and wild tomato evolutionary trees. The 12 wild

relative species also followed several nomenclature changes. One recent nomenclature with the ecological characteristics of the species was presented in Bauchet and Causse (2012) that compiled data from Peralta et al. (2008); Moyle (2008); Grandillo et al. (2011).

## 2. Modern phylogeny and taxonomy of the Tomato clade

In the past 30 years, numerous studies performed marker-assisted analyses using different types of molecular markers to uncover the phylogenetic organization of the Solanum genus. The first marker study leaded by Palmer focused on chloroplast DNA in 1982 [cpDNA (Palmer and Zamir 1982)] and managed to separate the Peruvianum group from the Esculentum group and revealed *S. lycopersicoides* and *S. juglandifolium* as outgroup species. Following this example, a few studies improved the genus phylogeny using chloroplastic DNA (Bohs and Olmstead 1997; Olmstead and Palmer 1997; Olmstead et al. 1999), mtDNA (McClean and Hanson 1986), nuclear RFLPs (Miller and Tanksley 1990) and AFLPs (Spooner et al. 2005; Zuriaga et al. 2009). These studies could already untangle most of the *Solanum* genus, separate and order the current species groups in the *Solanum* section *Lycopersicum* (namely Hirsutum, Peruvianum, Arcanum and Esculentum). The sequence data of internal transcribed spacer region of rDNA (Marshall et al. 2001), the Granule-Bound Starch Synthase (GBSSI) genes (Peralta and Spooner 2001), and the two nuclear genes from Zuriega and colleagues (Zuriaga et al. 2009) brought confidence into the main species classification and confirmed the tomato species phylogeny within the *Solanum* genus. Using 14 expressed sequence tags **(**ESTs) Roselius et al. (2005), completed the marker analyses on wild tomato accessions by estimating population genetics parameters such as nucleotide polymorphism or recombination rate. The pioneer work of Peralta, Spooner and Knapp refined the taxonomy in the genera, notably by the combined use of morphologic data and molecular markers genotyping (Peralta and Spooner 2000; Spooner et al. 2005; and see Peralta et al. 2008 for the taxonomic monograph). From the many studies they conducted, the topology demonstrated the monophyletic origin of the *Solanum* genera, section *Lycopersicum,* composed of 13 species.

The reference genome availability unlocked HTS studies focusing on the wild species speciation event and on the whole tomato genus phylogeny. For the sake of genus phylogeny clarification, the whole transcriptomes of 13 wild tomato species revealed evidences of diversification fuelled by at least three sources of adaptive genetic variation being “post speciation hybridization, rapid accumulation of new mutations, and recruitment from ancestral variation” (Figure 1; Pease et al. 2016). Multi-locus sequences of two wild species (*S. peruvianum* and *S.*

*chilense*) were implemented in coalescent-based models to infer the evolutionary processes of speciation (Stadler 2008). Two additional wild species *S. arcanum and S. habrochaites* added power to Bayesian methods to decipher their speciation (Tellier et al. 2011; Böndel et al. 2015). The population genetic approaches using 14,043 SNPs on 46 samples of *S. peruvianum* untangled the species complex into 4 separate species: *S. peruvianum* sensu stricto, *S. corneliomulleri*, *S. huaylasense* and *S. arcanum* (Labate et al. 2014), clarifying the organization of the clade. However, the real number of species of the wild species of tomato remains debated according to the criteria being used.

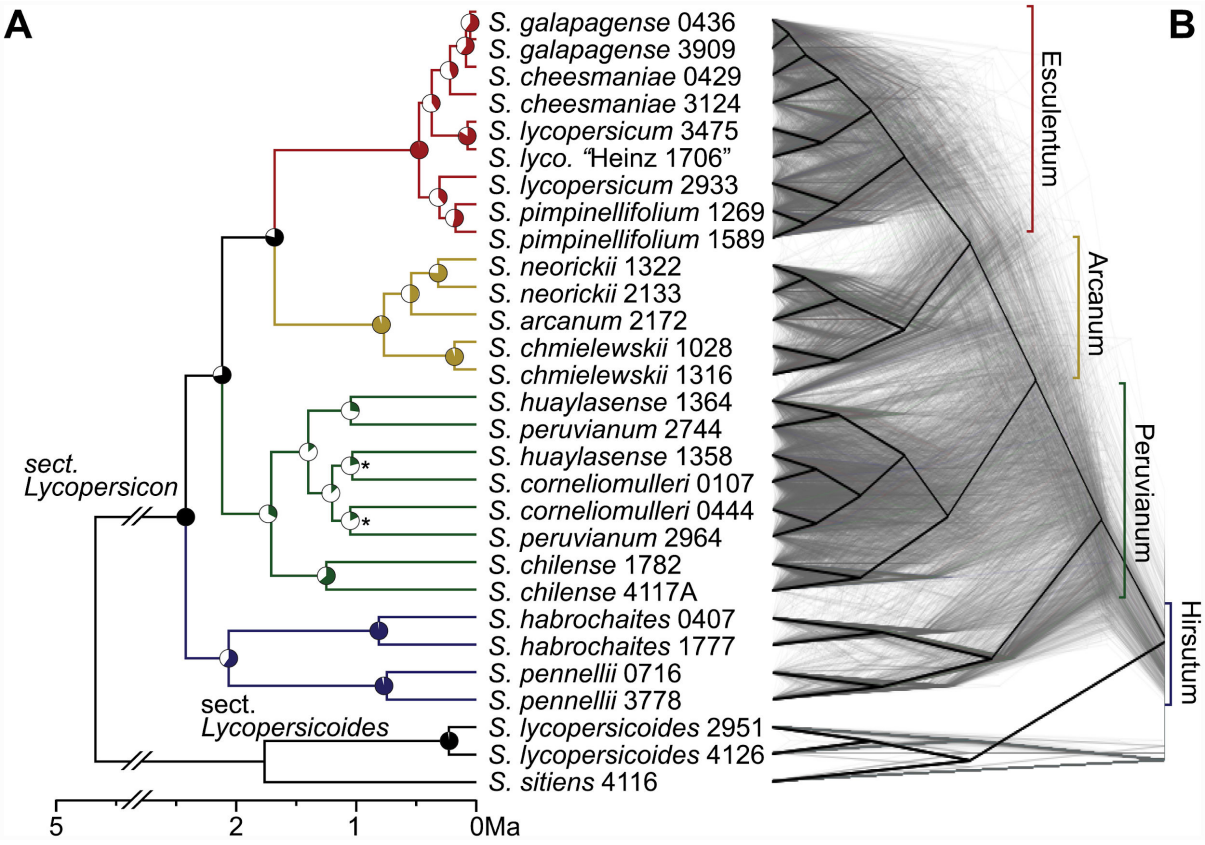



**Figure 1 (from Pease et al. 2016):** The phylogeny of Solanum sect. Lycopersicon. (A) A whole-transcriptome concatenated molecular clock phylogeny with section Lycopersicoides as the outgroup. Branch colours indicate the four major subgroups (labels on right). Pie charts on each node indicate majority rule extended bipartition support scores (out of 100) using trees from 100-kb genomic windows. All nodes are supported by 100 bootstrap replicates, except "*" denotes bootstrap support score of 68. (B) A "cloudogram" of 2,745 trees (grey) inferred from nonoverlapping 100-kb genomic windows. For contrast, the consensus phylogeny is shown in black.

The ease of closely related species hybridization within the tomato clade is the foundation of improvement of domesticated cultivars. Such hybridization is not quite current between *S. peruvianum* and *S. chilense* that are two distinct tomato species (Rick and Fobes 1975). The controlled hybridization with crop cultivars is a great opportunity to improve the domesticated tomato varieties.

## 3. Ecological Genomics of the tomato crop and its wild relatives

Ecological genomics aims at understanding the origin, history, and function of the observed natural biological variation, from nucleotide to community levels (Seehausen et al. 2014). In this context, the approach relies on ecological and genomic resources and provides an opportunity to precisely dissect genetic and developmental mechanisms, and to connect a genetic polymorphism to a phenotypic variation, as well as to directly demonstrate the ecological and evolutionary relevance of this phenotypic variation. Many of these studies have been performed in the wild tomato clade (*Solanum* section *Lycopersicon*), a group that has both exceptional diversity and genomic tools (see Haak et al. 2014, for a complete review). Within this section, we will focus on two major processes that are speciation and adaptation and report how much results did population genomics brought to these questions in the Solanum genus.

### *a. Speciation mechanism and reproduction barriers*

In tomato, the timing of speciation and the underlying molecular mechanisms of wild species divergence remained properly unresolved. Other nebulous scientific questions are still not resolved in this complex of species. The transition from self-incompatible to self-compatible reproduction system was partly induced by the domestication but the main molecular consequences are elusive. However, *S. habrochaites* is a wild self-compatible species of tomato revealing that the transition was independent. For example, genes involved in self-incompatibility are poorly characterized at the molecular levels (nucleotide diversity, gene expression levels).

As previously reported, strong reproductive barriers have been established between some of the species of the genera. Charles Rick's extensive work tested for these barriers by crossing all these species together during the 70 and 80's (Rick 1988; Rick and Chetelat 1995). As several speciation mechanisms seem to be at the origin of wild tomato diversification, there is a current debate on their respective roles/preponderance. On one side, traits responsible for prezygotic isolation (conferring ecological differentiation) are suspected to be the most important isolation barriers and most efficient in preventing gene flows between the species (Kirkpatrick and Ravigné 2002; Ramsey et al.

2003). On the other side, postzygotic barriers leading to hybrid unviability and sterility are more likely permanent and irreversible barriers to gene flow between species (Muller 1942; Coyne and Orr 2004). Moyle (2007) conducted a QTL mapping experiment to decipher the contribution of the pre- and postzygotic isolation between the plant species *S. lycopersicum* and *S. habrochaites* in a set of near-isogenic lines. They compared floral morphology between species and investigated sterility traits in hybrid crosses. However, the outcomes of this study remain limited as genome-wide associations were not evident: these traits showed a complex genetic architecture and association with centromeric regions warrant further fine-scale investigation, due to limited recombination. More recently, the role of the interspecific reproductive barriers (IRB) in limiting sympatric hybridization between closely related species was evaluated at three stages: prezygotic (floral morphology), post-mating prezygotic (pollen-tube growth), and postzygotic barriers (fruit and seed development) and were measured *in situ* in Peru by Baek et al. (2016). This study, based on 11 interspecific crosses demonstrated multiple IRB with three types of post-mating prezygotic IRB and strong postzygotic IRBs that prevented normal seed development by resulting from aborted endosperm and overgrown endothelium. However, hybridization was possible in some cases, notably from the pair *S. pennellii* × *S. corneliomulleri* with nearly developed seeds that produced viable F1 hybrids. In this latter case, molecular markers confirmed hybridity, which underlies the role of genomic tools for the study of this process. Thus, current studies on speciation mechanisms in wild tomatoes confirm the intricate role of pre- and postzygotic isolation and suggest that several scenarios underlie the speciation between two sister species.

From then on, population genomics revealed its potential to elucidate this question using RNASeq. Following Rick's investigations, extensive work focusing on postzygotic barriers has also been achieved in the species pair *S. peruvianum* × *S. chilense.* These two species are closely related with partly overlapping geographic ranges in northern Chile and southwestern Peru but are morphologically dissimilar. Roth (2017) demonstrated that crosses between these two species leaded to high proportions of non-viable seeds due to endosperm failure and arrested embryo development. On the basis of seed size differences in reciprocal hybrid crosses and developmental evidence implicating endosperm failure, they hypothesized that perturbed parental effects (*e.g.* genomic imprinting, or parent-specific allelic expression) were involved in the strong postzygotic barrier. They also conducted a transcriptomic screen in developing endosperms within intra- and inter-specific crosses and estimated the parent-of-origin–specific expression profiles using both homozygous and heterozygous nucleotide differences between parental individuals to identify

candidate imprinted genes (Florez-Rueda et al. 2016). As a result, they uncovered systematic shifts of “normal” (intraspecific) maternal:paternal transcript proportions in hybrid endosperms. Importantly, the genome-wide increased in maternal proportion almost entirely eliminated paternally expressed imprinted genes in *S. peruvianum* hybrid endosperm. Thus, they demonstrated that changes in parental expression proportions may be the underlying core process at play, leading to transcriptional regulation compromising the hybrid endosperm development and contributing to hybrid seed failure. However, at the opposite, they cannot reject that the transcriptional rewiring of the imprinted genes was the main source of perturbation of the essential developmental genes. Following this initial study, Roth (2017) extended this work with two additional species pairs and supported the common role of the genomic imprinting between nuclear and cellular endosperm types but also evidenced the genome-wide rewiring of gene expression and parental dosage in wild tomato hybrid endosperm as a major postzygotic barrier. More largely, these results are very interesting to reinvestigate the Endosperm Balance Number (EBN) hypothesis developed in the early 80’s in Capsella (Lagriffol and Monnier 1985). This hypothesis was proposed to explain the basis of normal seed development after intra and inter-specific crosses, through a 2:1 maternal to paternal ratio in the hybrid endosperm. Up to now, it was mostly not possible to properly test for how EBN may act as powerful isolating mechanism (Carputo et al. 1999).

The release of Recombinant Inbred Lines (RIL), linkage maps and the genome of the domesticated tomato (*Solanum lycopersicum*) were valuable tools for the genetic analysis of interspecific reproductive barriers. It provided the basis for QTL detection, read mapping, gene annotation, shedding light on the underlying mechanisms involved in reproductive barriers in the tomato genera. However, transgenic methodologies are new tools that are providing opportunities to test the candidate loci involved in these barriers, while the complementation of proteomic and transcriptomic offers insights into the molecular regulation of gene expression to provide a clearer picture of the interspecific reproductive barriers present in wild tomato relatives through the identification of new candidate genes or proteins (Bedinger et al. 2011). Finally, Li and Chetelat (2010, 2015) deciphered the Unilateral interspecific Incompatibility (UI) system and identified two genes that block cross-hybridization between related species, typically when the pollen donor is self-compatible and the pistil parent is self-incompatible (SI): ui1.1, a pollen UI factor in tomato, which encodes an S-locus F-box protein and ui6.1, which encodes a Cullin1 protein that functions in both UI and SI.

### b. *Ecological Adaptation*

Darwin proposed that phenotypic differentiation among populations resulted from differential adaptation in response to environmental heterogeneity (Kawecki and Ebert 2004). This mechanism is relatively frequent and has been proven experimentally by connecting physiological, genetic and ecological data to measure the fitness over evolutionary timescales. Two main factors have been proposed to influence ecological adaptation: abiotic and biotic stresses, which may also interact together. The advent of high-throughput genomics allowed refining our knowledge of the adaptation mechanisms. The tomato genus was extensively used towards this objective notably thanks to its large geographic range (Figure 2). These contrasting environments are characterized by different stress conditions such as drought, salt, cold and heat. Hereafter, we describe a limited number of uses of genomics to document the molecular mechanisms of adaptation to abiotic stress in crop and wild tomatoes. For a complete review, including adaptation to biotic stress, see Haak et al. (2014).

Among wild tomatoes, it has been demonstrated that the greatest axes of differentiation between species are average annual rainfall and temperature (Nakazato et al. 2010). QTL mapping experiments reported that both *S. chilense* and *S. pennellii* developed distinct strategies to adapt to drought stress. In addition, the comparison with the domesticated tomato identified QTL associated with eco-physiological trait variation and identified a various and complex genetic architecture based on both main effect and transgressive QTLs (Muir and Moyle 2009). Contrasted patterns of nucleotide diversity patterns of local adaptation at drought related candidate genes in wild tomatoes (*S. peruvianum* and *S. chilense*), identified at two major loci in the abscisic acid signalling pathway, were observed. On one side, *LeNCED1* exhibited very low nucleotide diversity relative to the eight neutral reference loci that were surveyed in populations of these two species. This suggested that strong purifying selection has been acting on this gene. On the other side, *pLC30-15* exhibited higher levels of nucleotide diversity. Additionally, for these two loci, in particular in *S. chilense*, higher genetic differentiation (Fst) between populations than for the reference loci, indicated local adaptation and in the more drought-tolerant species *S. chilense*, one population (from Quicacha) showed a significant haplotype structure, which appeared to be the result of positive (diversifying) selection (Xia et al. 2010).

Local adaptation is crucial when a species colonizes new habitats. The tomato wild relative species *S. chilense* is an example of native range expansion in southern America from North to South.

It provides a strong experimental framework to test for differential hypothesis underlying the mechanisms of local adaptation through colonization. Böndel et al. (2015) tested whether local adaptation occurred more frequently in large ancestral populations or in small derived populations using a population genomic approach. They conducted a population genetic analysis and inferred the past demography of *S. chilense* populations on pooled-sequencing data from 30 genes (8,080 SNPs). Across Chile and Peru, 23 *S. chilense* populations were sampled according to the north to south colonization. Along this cline, a decrease of genetic variation was associated with a relaxed purifying selection and an increasing proportion of non-synonymous polymorphism from the study of the distribution of fitness effect, and by population substructure with at least four genetic groups. In other words, the north to south cline is associated with an increase in deleterious mutations, potentially conferring a decreased adaptive potential to southern populations. Patterns of population structure, natural selection, and linkage disequilibrium within these *S. chilense* populations confirmed previously inferred population-specific demographic histories (Arunyawat et al. 2007).

Similarly, spatial genetic analyses revealed clinal pattern for other wild tomato species such as the wild relative *S. peruvianum* and *S. pimpinellifolium* and the cultivated *S. lycopersicum* (Nakazato and Housworth 2011; Nakazato et al. 2012) and in the related Solanaceae species *S. lycopersicoides* and *S. sitiens* (Albrecht et al. 2010), which occur in sympatry with *S. chilense* in northern Chile (Peralta et al. 2008). These patterns of clinal variation of nucleotide diversity were correlated to seed bank size. The combination of ecological and genomic data provided evidence and putative parameters for seed bank in both *S. chilense* and *S. peruvianum* (Tellier et al. 2011). In this study, the inferred difference in germination rate between these two species reflected divergent strategy of adaptation for seed dormancy, that agreed with previous population genetic analyses and the ecology of these two-sister species. Overall, the ‘seeds‘ strategy relied on spending on average, a shorter time in the soil in the specialist species (*S. chilense*) than in the generalist species (*S. peruvianum*).

Using whole transcriptomes from the 13 species of the Solanum genera, Pease et al. (2016) used population genomics and not only identified the ecological and genetic factors that promoted the species radiations and inferred the species phylogeny (see Part I), but they also found evidence for at least three sources of adaptive genetic variation that fuel species radiations. First, they detected introgression events between the early-branching lineages and more recently between individual populations. This supported the hypothesis of adaptive benefits through hybridization.

Second, they evidenced lineage-specific *de novo* evolution for loci involved in the production of red fruit colour. Third, they detected environment-specific sorting of ancestral variation among populations that come from different species that shared common environmental conditions. Overall, these results indicated that multiple genetic sources can promote a rapid diversification and endow the speciation mechanism in response to ecological adaptation. Last but not least, this study highlighted the complexity of both ancient and recent species radiations, using a combination of ecological and genomic data.

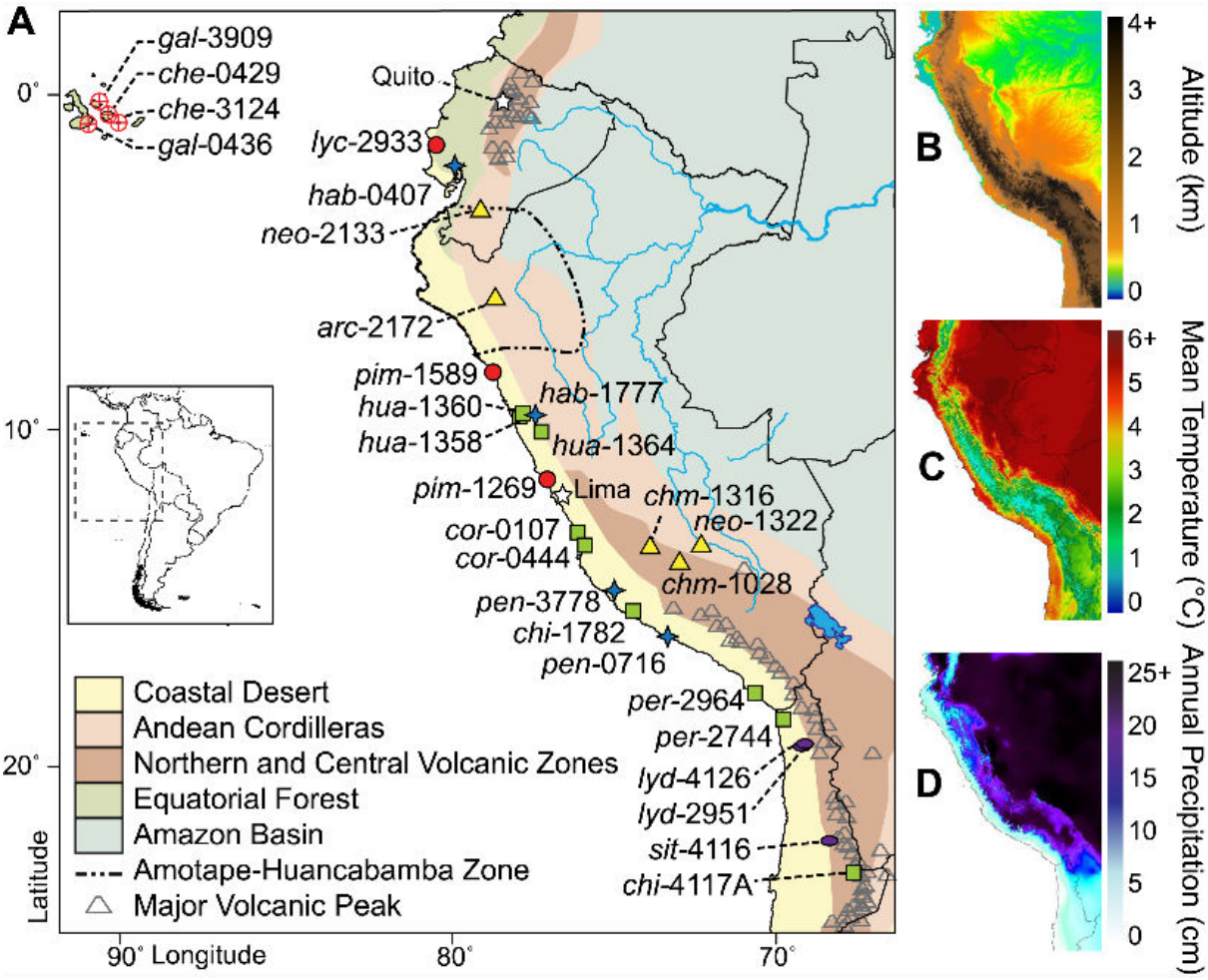



**Figure 2 (from Pease et al. 2016):** Geographic distribution and ecological diversity of sampled populations of wild tomato. (A) Wild tomato species inhabit diverse ecological zones (shaded regions) along the western coast of South America and the Galápagos Islands. For each sample location, labels indicate species and accession number, and symbols denote major phylogenetic groupings (circle = Esculentum, triangle = Arcanum, square = Peruvianum, star = Hirsutum, oval = outgroup; base map modified from original from http://www. freevectormaps.com). High variation of (B) altitude, (C) mean annual temperature (D), and annual precipitation across the habitat range of wild tomato species (data from http://www.worldclim.org; plotted using GRASS GIS http://grass.osgeo.org/).

In brief, the tomato genera, that includes 12 wild species covering a large geoclimatic range is an excellent framework to investigate the origin and history of the biological variation that occurs at the phenotypic and genomic levels (Haak et al. 2014). High-throughput genomics extends our understanding of these past processes. The combination of the 'omics' approaches, notably transcriptomics with metabolomics and proteomics, provides an exceptional opportunity to get clearer interpretation of the forces at play in the processes of speciation and adaptation within the Solanum genus but also in sister genus such as Capsicum. The availability of a high-quality genome sequence of cultivated tomato was key towards these amounts of results, but efforts should be brought towards a high-quality reference genome for each of the 12 wild species. The use of third generation sequencing technologies (i.e. Oxford Nanopore sequencing) is about to deliver such promises.

## 4. Genomic footprints of domestication and modern breeding stages

### *a. Deciphering the domestication and breeding history*

Comparative genomics has proven to be a valuable tool to decipher evolutionary mechanisms and forces that occurred over macro and micro timescales. Comparing patterns of nucleotide patterns is the basic idea behind this approach to highlight constrained loci by evolutionary forces. Both domestication and modern breeding stage (also called 'improvement') are appropriate models for studying adaptation, genome evolution, and the genetics and evolution of complex traits. For example, the accumulation of non-synonymous variants (i.e. the so-called 'genetic cost of domestication', see Lu et al. (2006) or the original hypothesis and Moyers et al. (2018) for an updated review, and selective sweeps (stretch of homozygosity due to breeding practises) were evidenced between crop and wild accessions in many crop species such as soybean (Lam et al. 2010), maize (Hufford et al. 2013) or rice (Xu et al. 2012). Comparative expression profiling extended the approach in a few crops, such as maize (Swanson-Wagner et al. 2012; Lemmon et al. 2014), cotton (Rapp et al. 2010) or common bean (Bellucci et al. 2014). In tomato, the consequences of the domestication syndrome have been deeply studied for phenotypic traits such as growth habit (plant vigour and flowering time) and fruit traits (set, size, shape, colour and morphology) and many major genes and QTLs have been identified during the last decades (Grandillo and Tanksley 1996; Doganlar et al. 2000; Tanksley 2004; Bai and Lindhout 2007; Chakrabarti et al. 2013). The use of 'omics' (i.e. HTS) also shed light onto the genomic footprints of domestication and modern breeding in the tomato.

Using genomics, tomato domestication was clarified, notably by delineating the position of *S. l. cerasiforme* and its role in this process. To do so, a very large collection of >1,000 accessions was screened using the SOLCAP SNP array (>8,000 SNPs). Tomato domestication seems to have followed a two step-process; a first domestication in South America and a second step in Mesoamerica (Blanca et al. 2015). The distribution of fruit weight and shape alleles supported that domestication of *S. cerasiforme* occurred in the Andean region and clarified the biological status of this genetic group as a true phylogenetic group within tomato.

### *b. Variation of nucleotide diversity patterns*

The strong human selection induced by domestication and later on by crop improvement, left footprints on the plant genome that can be tracked through the genome-wide study of nucleotide diversity with summary statistics such as $\pi$ and Tajima's D. Then, from these summary statistics, selective sweeps or genetic bottlenecks can be evidenced. In tomato, the genome-wide reduction in nucleotide diversity has been one of the most obvious genetic mark of such bottlenecks during the domestication of *S. lycopersicum* from its closest wild relative species *S. pimpinellifolium*. Miller and Tanksley (1990) reported that the amount of genetic variation in the SI species (i.e. *S. peruvianum*) far exceeded (-95%) that found in SC species (*S. lycopersicum*) from the analysis of RFLP markers. More recently, this loss has been supported but revised by many studies (The Tomato Genome Consortium 2012; Koenig et al. 2013; Lin et al. 2014; Blanca et al. 2015; Sauvage et al. 2017; Sahu and Chattopadhyay 2017). We observed variable but drastic reduction of the total nucleotide diversity ($\frac{\pi CROP}{\pi WILD}$= 0.37 reported in Lin et al. (2014)from the comparison between *S. lycopersicum* and *S. pimpinellifolium* at the genome-wide scale and $\frac{\pi CROP}{\pi WILD}$= 0.65 - reported in Sauvage et al. 2017 from the comparison between *S. lycopersicum* and *S. pimpinellifolium* at the transcriptome-wide scale). However, this drastic reduction has to be cautiously interpreted because these average estimates across the genome may not reflect specific genomic regions.

The extensive use of wild germplasm for breeding purpose was a common practice during the improvement stage in tomato. This had an impact on genome structure/architecture as shown by the extensive work achieved by Labate and collaborators (2009). When examining genome-wide patterns of nucleotide diversity, small chromosomal regions show non-randomly distributed regions of higher nucleotide diversity in cultivated compared to wild accessions. In *S. lycopersicum*, these regions showed increased allele sharing with *S. pimpinellifolium*, indicating recent introgressions

from this species or a closely related other one. Koenig et al. (2013) defined 550 candidate introgressed genes in the reference genome of Heinz1706 and 2,479 in the cultivated accession M82. The large number of candidate loci introgressed in M82 highlights the challenge of linkage drag during breeding using wild accessions, and may contribute to reduce genome-wide divergence in nucleotide sequence between cultivated and wild accessions. Similar observations were reported in Blanca et al. (2015) when comparing contemporary *S. lycopersicum* to vintage accessions and in Sauvage et al. (2017) when comparing *S. lycopersicum* to *S. pimpinellifolium*, especially on chromosome 9. Additionally, evidences of strong genetic bottleneck and relaxation of purifying selection were reported. Estimates of dN/dS in *S. lycopersicum* supported the accumulation of potentially deleterious mutations during its cultivation (Koenig et al. 2013). In contrast, Sahu and Chattopadhyay (2017) identified a continuous and strong purifying selection in the cultivated tomato which may be required to maintain some favoured agronomic trait. About 1% (8.76 Mb) of the tomato genome (distributed across seven chromosomes) showed very strong purifying selection with Tajima's D estimates lower than −3.0. Breeding may have also contributed to fix haplotypes and reduce nucleotide diversity by favouring one allele of interest (i.e. hard selective sweep). A total of 186 domestication sweeps ($\frac{\pi\ S.cerasiforme}{\pi\ S.pimpinellifolium}$) and 133 improvement sweeps ($\frac{\pi\ S.cerasiforme}{\pi\ S.lycopersicum}$) covering nearly 8.3% (64.6 Mb) and 7.0% (54.5 Mb) of the species genome were identified, witnessing the frequency of allele fixation during the history of tomato breeding (Lin et al. 2014). Overall, both domestication and improvement sweeps overlapped with known QTL, notably related to fruit weight, a major trait affected during these two stages of the tomato history (i.e. locus *fw2.2, fw3.2...*).

### *c. Domestication and modern breeding induced a transcriptome rewiring*

The comparative genomics approach was extended by using gene expression levels to decipher the genome-wide transcriptional changes induced during the domestication and improvement stages of the tomato history. Expression and co-expression patterns were investigated and showed that specialized as well as general pathways have been affected during both stages. Itkin et al. (2013) showed how tomato turned from "nasty to tasty". More precisely, metabolic pathways and genes directing the synthesis of some anti-nutritional compounds (i.e. Steroidal GlycoAlkaloids - SGAs) in potato and tomato were elucidated. Comparative co-expression analyses between tomato and potato coupled with chemical profiling revealed ten genes partaking in SGA biosynthesis. Six of them form a cluster on chromosome 7, whereas an additional two are adjacent in a duplicated

genomic region on chromosome 12. The Silencing *GLYCOALKALOID METABOLISM 4* pathway prevented accumulation of SGAs in tomato fruit and in potato tubers. This demonstrated that domestication down-regulated entire specialized metabolic pathways, locking the production of antinutritional compounds.

Patterns of differential expression and co-expression between cultivated and wild tomato species showed major transcriptional changes in genes related to stress response, defence response, photosynthesis, response to high light, and redox pathways (Koenig et al. 2013). These molecular functions partly overlapped with genes related to response to stress, the generation of precursor metabolites and energy, metabolic process, the epigenetic regulation of gene expression, and carbohydrate metabolism additionally identified in Sauvage et al. (2017). Enrichment for these categories indicated that abiotic and biotic stresses have played a major role driving transcriptional variation along tomato history. In addition, the comparison of genomic and transcriptomic patterns conducted in Sauvage et al. (2017), showed that both synonymous and non-synonymous polymorphism rates tended to be higher in the wild group than the cultivated group. This trend was significantly more pronounced for differentially expressed genes (DEG) between crop and wild tomato accessions, than for the non-differentially expressed ones, indicating that purifying selection was significantly weaker in DEG compared with non-DEG. Altogether, this suggests that purifying selection tends to be stronger among DEG in the wild genetic group.

## Part III: Population genomics to sustain modern breeding

There are two strong interests in studying the crop wild relative species such as wild tomatoes: (I) Use the wild relative species to better understand processes and modification triggered by domestication into crop plants (Abbo et al. 2014) and (II) identify and introgress wild relative genes of interest to gain new genetic diversity following the strong diversity bottlenecks and thus increase the crop fitness (Ohmori et al. 1995, 1998). Since the pioneer work of Steve Tanksley's research group, molecular markers were used to construct a high-density genetic map of the tomato genome (Tanksley et al. 1992) and dissect quantitative traits into Mendelian factors or QTL (Quantitative Trait Loci) (Paterson et al. 1988; Tanksley et al. 1992). This also allowed to positionally clone the genetic factors underlying major mutations or quantitative traits (Paterson et al. 1991). The low polymorphism detected by RFLP and PCR markers compelled geneticists to study interspecific segregating populations, which were more polymorphic. This also underlined the interest of the wild

relative species as a source of new diversity. With the availability of SNP markers, it became possible to study large collections of varieties, develop GWAS and advance our knowledge about the genome structure such as linkage disequilibrium decay, the structure of haplotypes, the distribution of recombination and to identify early introgressions from wild species.

## 1. Introgressions from crop wild relative species improved the crop tomato

The crucial role of crop wild relatives has been identified for many crops (Vincent et al. 2013; Brozynska et al. 2015), but it is particularly pronounced for tomato breeding. This was already suggested by the pioneer work of Charles Rick (1990)who showed the existence of several disease resistances in wild tomato species. More than 200 pathogens infect the crop tomato (Bai and Lindhout 2007). Heirloom varieties are usually susceptible to all of them. Thus, wild relatives were first screened for disease resistances and many monogenic dominant genes were discovered. They were subsequently introgressed into cultivars and nowadays modern hybrids carry up to eight disease resistance genes. The introgression required the identification of molecular markers linked to these genes and many of them are now located on the genome (Causse and Grandillo 2016). Following the mapping effort, tomato was used as a model species to clone these genes and decipher their structure and their molecular organisation (Martin et al. 1993). Wild germplasm has played a crucial role in the modern breeding of cultivated tomato (Stevens and Rick 1986), triggering interest for wild tomatoes species and for the evolution of the group as a whole (Labate et al. 2007).

During the sequencing of the tomato reference genome, the introgression of several chromosomal segments related to *S. pimpinellifolium* was shown (TGC 2012). These introgressions, probably due to the first introgressions of disease resistance genes were detectable on several chromosomes, suggesting several rounds of introgression.

The large size of introgressions from wild relative species was first shown by (Young and Tanksley 1989). This was confirmed at the genome scale by Lin et al. (2014) who detected in a set of modern F1 hybrids a large exotic fragment on chromosome 9 (more than 50 Mb in length) carrying the tobacco mosaic virus resistance gene *Tm-2$^a$* derived from *S. peruvianum*. In addition, they detected two other major introgressions on chromosome 6: one (>25 Mb in length) carrying the root knot nematode resistance gene *Mi-1* introgressed from *S. peruvianum* and the other (+30 Mb in length) carrying the tomato yellow leaf curl virus resistance gene *Ty-1* from *S. chilense*. Even after many generations of backcrossing, these introgressed fragments remain intact, possibly due to

chromosomal rearrangements or a centromeric location that would inhibit recombination, as in the case of *Ty-1* and *Mi-1* (Seah et al. 2004; Verlaan et al. 2013).

## 2. Dissecting the genetic architecture of agronomical traits

Quantitative trait mapping revealed the potential of crop wild relatives even for un-targeted traits. Due to the low genetic diversity within the cultivated compartment (Miller and Tanksley 1990), most of the first mapping populations were based on interspecific crosses between a cultivar and a related wild accession from the Lycopersicon section (as reviewed by Foolad (2007); Labate et al. (2007); Grandillo et al. (2011)) or from *Lycopersicoides* (Pertuzé et al. 2002) and *Juglandifolia* group (Albrecht et al. 2010). However, intraspecific crosses, notably with cherry tomatoes have proved their interest notably on fruit quality aspects (Saliba-Colombani et al. 2001). All those populations allowed discovering and/or characterizing a myriad of major genes and QTLs involved in various traits (recent synthesis in Grandillo and Cammareri 2016).

Introgression Lines (IL) derived from interspecific crosses allowed dissecting the effect of unique chromosome fragments from a donor (usually a wild relative species) introgressed into a recurrent elite line. ILs were used for fine mapping and positional cloning of several genes and QTL of interest. The first IL library was developed between *S. pennellii* and *S. lycopersicum* (Eshed and Zamir 1995; Zamir 2001). This progeny was used to identify QTLs for fruit traits (Causse et al. 2004), anti-oxidants (Rousseaux et al. 2005), vitamin C (Stevens et al. 2007) and volatile aromas (Tadmor et al. 2002). QTL mapping power was increased compared to biallelic QTL mapping population, and was again improved by the constitution of sub-IL set with smaller introgressed fragments (Ofner et al. 2016). Such exotic libraries were thus designed with several species, involving *S. pimpinellifolium* (Doganlar et al. 2002b), *S. habrochaites* (Monforte and Tanksley 2000; Finkers et al. 2007) and *S. lycopersicoides* (Canady et al. 2005). Introgression lines were also used to dissect the genetic basis of heterosis (Eshed and Zamir 1995). Heterosis refers to a phenomenon where hybrids between distant varieties or crosses between related species exhibit greater biomass, speed of development, and fertility than both parents (Birchler et al. 2010). Heterosis involves genome–wide dominance complementation and inheritance model such as locus–specific overdominance (Lippman et al. 2007). The potential of related wild species even for improving unexpected traits was shown as, for instance, some QTL alleles increasing the red colour of the fruit were discovered in *S. pennellii,* a green-fruited species (Causse et al. 2004). Interesting alleles at QTL for fruit volatiles were also detected in several interspecific progenies (Klee 2010).

## 3. Molecular bases of trait diversification

Tomato domestication and later diversification of fruit types, led to a large morphological diversity in tomato fruit (with small to large, round, blocky, elongated, pear shaped fruits, with colour ranging from red to green, white, black, pink, orange or yellow). On the contrary, wild tomato species carry small, round red or green fruits, with a limited intraspecific phenotypic diversity. Using molecular markers, the genetic control of fruit traits has been widely dissected (Grandillo et al. 1999; Lippman and Tanksley 2001; Barrero and Tanksley 2004). The first QTL controlling fruit weight variation, fw2.2, was cloned (Frary 2000) followed by several mutations/QTL involved in fruit shape: LC and FAS which increase locule number and fruit size (Cong et al. 2008; Muños et al. 2011), OVATE which gives ovoid fruit shape (Liu et al. 2002) and SUN which gives an elongated fruit shape or the oxheart shape when associated to LC and FAS (van der Knaap et al. 2002; Xiao et al. 2008). It was then shown that the combination of alleles at these four genes were responsible of most of the diversity of fruit shape present in cultivated germplasm (Rodriguez et al. 2011). The allelic distribution of the four genes was then associated with morphologic, geographical and historical data in a collection of diverse cultivated accessions and a model for fruit shape evolution in tomato was established suggesting that selection occurred in distinct chronologic and historic periods: LC arose first, followed by OVATE, both in *S.l. cerasiforme* background but in distinct populations. FAS arose later in a LC background. The presence of these three mutations in Latin American germplasm suggested Pre-Columbian mutations. Combined with fw2.2, they must have strongly contributed to the increase in fruit size during tomato domestication. On the contrary, SUN mutation is not carried by any Latin American material tested, suggesting that SUN mutation appeared post domestication in European material (probably in Italy). This study also showed that the selection for fruit shape is strongly responsible for the underlying genetic structure in tomato cultivars.

## 4. Breeding shaped the genetic structure of modern cultivars

As previously stated, selection during domestication and subsequent breeding considerably reduced the genetic diversity of the tomato crop. To further improve the crop, segments of wild tomato genomes were introgressed into modern cultivars (Rick 1960). To better understand how modern breeding had changed the tomato genome, Sim et al. (2011) studied the population structure of 70 tomato lines (with 173 markers) and found clusters that separated the cultivated tomato into processing, fresh-market, vintage and landrace varieties. A similar study detected a longer linkage disequilibrium decay in processing tomatoes (7 to 14 cM) and in fresh market tomato (3 to 16 cM)

than in vintage cultivars (6 to 8 cM), which revealed the strong selection cost that modern breeding induced in processing and fresh market tomato varieties (Robbins et al. 2011).

More recently, Lin et al. (2014), sequenced 360 tomato genomes: 333 representing the diversity of types and varieties from the red-fruited clade (*S. pimpinellifolium*, *S. lycopersicum* var. *cerasiforme* and *S. lycopersicum*) including 166 big-fruited *S. lycopersicum*. They detected two main groups in *S. lycopersicum*: the first including accessions of *S. lycopersicum* with big fruits paired with the Non-South American S. *lycopersicum* var. *cerasiforme* and the second composed by the S. *lycopersicum* var. *cerasiforme* that were originated from south America. With a higher resolution (K=4) they could as well detect the processing tomato cluster. They focused on PIM (*S. pimpinellifolium*), CER (S. *lycopersicum* var. *cerasiforme*) and BIG (*S. lycopersicum*) clusters and observed domestication induced diversity decrease by measuring the number of sites that were polymorphic in each group, from the close wild relative PIM (30.4% of the total 3.5 million SNPs) to the BIG (2.8%) with the intermediate group of CER (6.6%). Following this polymorphism detection, they showed a strong difference in linkage disequilibrium decay occurring between SNPs at physical distance of 8.8 kb in PIM, 256.8 kb in CER and 865.7 kb in BIG (figure 3). The domestication and improvement swept regions occupied nearly 25% of the assembled genome, these 25% of sweeps experience a strong LD, costs of domestication, and will be limiting for future conventional tomato improvement.

The 1,008 tomato accessions that were genotyped using 7,720 SNPs by Blanca et al. (2015)completed the previous analyses. In this study, the heterozygosity expected and observed were higher in PIM (He = 0.21 / Ho = 0.042) than in CER (He = 0.17 / Ho = 0.023) and in BIG (He = 0.12 / Ho = 0.012). Known introgressions were detected in modern cultivar compared to so-called vintage ones, by measuring a higher heterozygosity due to the wild introgressions (He = 0.12 vs. 0.09). In a recent paper, Sahu and Chattopadhyay (2017), detected 2,439 SNPs that were only polymorphic in wild accessions, these wild variants being part of 1,594 genes (868 SNPs were located up- and downstream of these genes). With this study they confirmed that chromosomes which were the most affected by domestication and presented high diversity loss were the chromosomes 1, 2, 3, 8 and 10. These chromosomes are including the chromosome 2 that is known since 1964 (Kerr and Bailey 1964)as bearing three of the few genes responsible for the fruit shape and size (*LC, fw2.2* and *Ovate*).

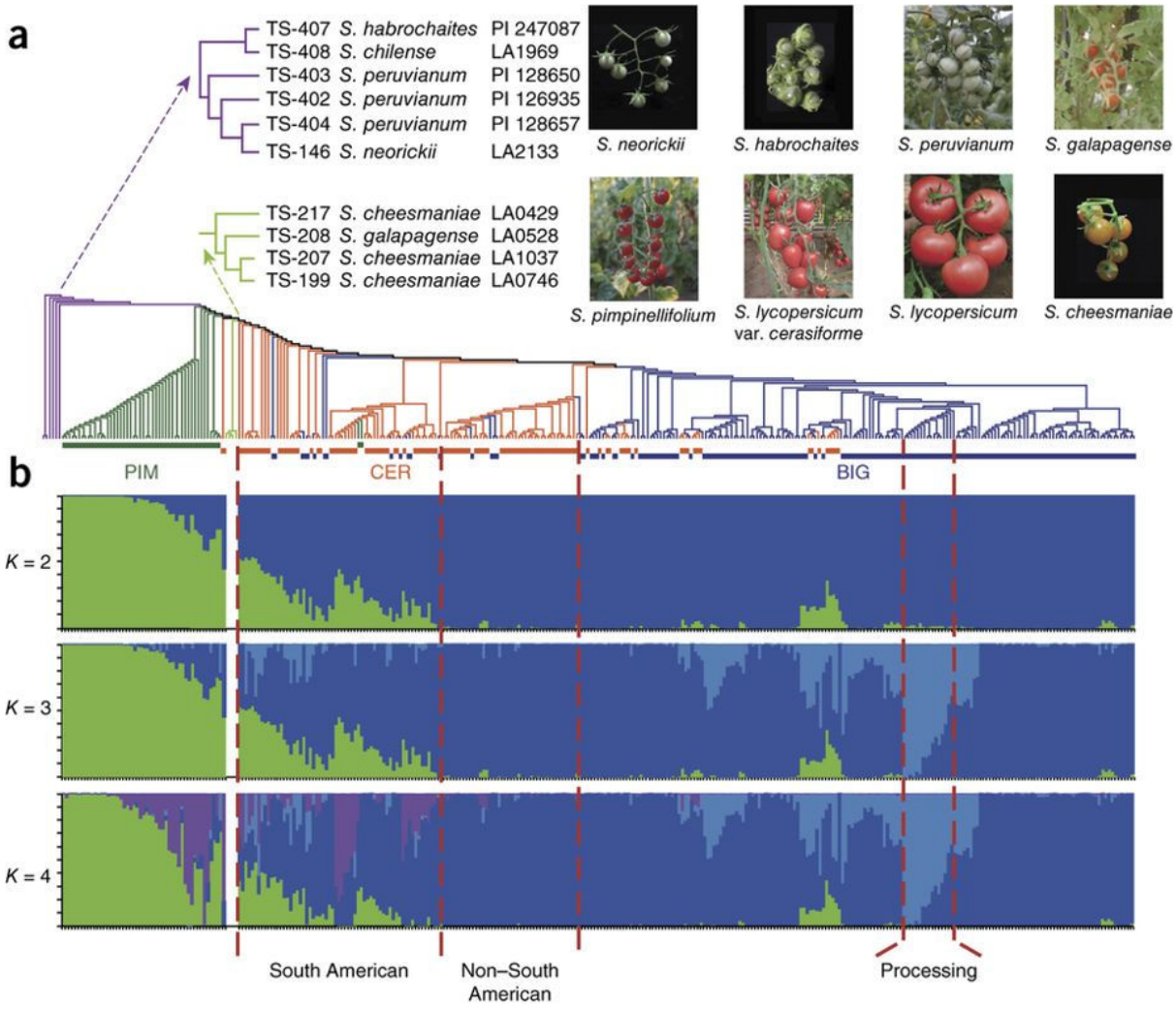



**Figure 3.** (from Lin et al. 2014): (a) The neighbour-joining tree of the population (331 accessions from the red-fruited clade and 10 wild accessions) was generated using 20,111 SNPs at fourfold-degenerate sites. The bars indicate the PIM (green), CER (orange) and BIG (blue) lines. The two branches containing wild accessions are enlarged for visualization. Typical fruits of the species studied are shown. (b) Model-based clustering analysis with different numbers of clusters (*K* = 2, 3 and 4). The *y* axis quantifies cluster membership, and the *x* axis lists the different accessions. The orders and positions of these accessions on the *x* axis are consistent with those for the neighbour-joining tree. South American CER, non–South American CER and processing tomato clusters are separated by dashed red lines.

### 5. Genome-wide association approach extended the knowledge of the genetic architecture of agronomical traits

In plants, the QTL approach has been largely used in biparental and multi-parental crosses (i.e. MAGIC - Laura et al. (2014) - or NAM populations). However, this approach is restricted in allelic diversity limiting the genomic resolution to map genetic determinants (Borevitz and Nordborg 2003).

The genome-wide association approach (GWAS) was proposed to overcome the main limitations of traditional gene mapping by (i) providing higher resolution using ancestral polymorphism at the population level and (ii) using panels of individuals from populations in which commonly occurring genetic variations can be associated with phenotypic variation. The availability of high-density SNP arrays (Sim et al. 2012; Víquez-Zamora et al. 2013) and sequencing data allowed genome-wide scans to test for significant associations between molecular markers and the quantitative trait variation. While firstly applied in large studies of human disease, that successfully identified candidate loci (Hindorff et al. 2009), GWAS was adopted in plants only a decade ago. Overall, these successful studies identified loci that explain large portions of phenotypic variation (Brachi et al. 2011).

In major crop species, GWAS was applied to decipher the genetic architecture of complex quantitative traits and benefited from statistical and technical developments. More precisely, the implementation of mixed linear models (MLM) to account for population structure and kinship, estimated in the studied panel, allowed detecting associations with a higher accuracy. Similarly, correction for multiple testing (i.e. FDR or Bonferroni corrections) removed false positive associations sorting out the most promising candidate loci. Additionally, the size of the GWAS datasets in major crops followed the trend of the power of high-throughput genotyping and sequencing technologies. From few SSR or SNP makers, a decade ago, actual genomic datasets rely on full length genome sequence for hundreds of individuals. Tomato was not an exception with numerous GWA studies conducted during the last decade, notably for agronomic traits such as fruit morphology, metabolomic content or genotype by environment interactions (GxE). In more details, the first association studies investigated fruit quality using limited sets of SNP (<100) spread over the chromosome 2 (Ranc et al. 2012). Then, rapidly, with the development of the SOLCAP SNP genotyping array (nearly 8000 SNPs), genome-wide level GWA experiment were conducted to decipher the genetic basis of agronomical traits such as fruit morphology or fruit metabolite contents (Sauvage et al. 2014; Ruggieri et al. 2014; Sacco et al. 2015; Zhang et al. 2015; Bauchet et al. 2017a, b). Then, low coverage sequencing and full genome sequencing provided a broader coverage of the tomato genome, increasing the power to detect new associations notably related to fruit colour (Lin et al. 2014), agronomical traits (Shirasawa et al. 2013; Ye et al. 2017), flavour components (Tieman et al. 2017) or extensive sets of primary and secondary metabolites (Zhu et al. 2018). However, within these latter studies, the interactions between the genotype and its surrounding environment were not considered until Albert et al. (2016) provided a GWAS study of the impact of drought stress onto agronomical and fruit quality traits in tomato, opening the door to further GxE experiments, notably

related to biotic and abiotic stress tolerance. Overall, these studies made an extensive use of the combination of population genomics and germplasm collection. They deepened our knowledge on tomato genome dynamics in terms of recombination patterns (through the study of the LD decay), identified candidate loci that were functionally validated, proving the validity of the approach in this species and more largely in plants.

However, in tomato, as in major crop species, limitations to sustain the discovery of new candidate loci underlying complex traits remain. Breakthroughs have been made in the field of high-throughput phenotyping, such as nano-sensors assisted phenotyping (Dalal et al. 2017) which complement the production of population genomics data in this field of research. These latter technologies would be easily transferred to decipher the genetic architecture of local adaptation processes, for example, by providing large amounts of data for a reasonable cost. Another limitation is the statistical correction applied for multiple testing that inherently lower the power of the association approach towards low to medium effect loci. At this stage, population genomics will be of great help to tackle this problem. Haplotype determination methods are more mature procedures, thanks to the HapMap human project. However, these procedures have been sparsely applied in crop species (Wang et al. 2013). Reports in maize, rice or soybean demonstrated the power of the approach for adaptive traits such as flowering time (Van Inghelandt et al. 2012), sugar metabolism (Lestari et al. 2011) or salinity resistance (Patil et al. 2016), respectively. Besides the identification of promising candidate loci, the use of haplotypes provided further knowledge of the demographic or selective history of these loci. The same approach can be applied in tomato that experienced drastic changes in nucleotide diversity patterns along its domestication and modern breeding phases. Thus, haplotype makers will strengthen biological interpretations obtained from quantitative genetics and population genomics, offering a broader view of selective forces that acted on loci related to traits of agronomical interest for which the molecular determinants have been identified by GWAS. Dealing with the missing heritability is another limitation of the GWA approach (Brachi et al. 2011), that limitation could be unveiled by population genomics based on the analysis of epi-markers. The approach was successfully applied, notably in human for common diseases (Rakyan et al. 2011) as epigenetic variation affects genes function and can contribute to common disease, and, in *Arabidopsis thaliana*, for local adaptation (Dubin et al. 2015). Overall, there is a unique opportunity to merge population genetics and population genomics to get the best of both worlds in sustaining breeding efforts while deciphering the selective history of agronomical loci in crops, such as tomato.

## Part IV: Prospects for future research

### a. Towards a pan-genome in tomato

Large scale genomic characterization of genetic diversity in plants is already ongoing, especially with the re-sequencing of large sets of accessions. In 2018, over five hundred genome sequences are publicly available in tomato and future projects aim to sequence up to thousands of accessions. These data allowed the identification of domestication footprints and track hybridization events, for example (Lin et al. 2014). Mining and leveraging the sequence data in such large-scale projects require a pan-genomic approach. A pan-genome structure that describes the full complement of genes in a single species, has multiple advantages over a single, linear reference genome sequence for population genomics and plant breeding applications. The approach was applied in crop and wild accessions of rice. Identifying conserved and variable regions allowed to pinpoint new causal variants that underlie complex evolutionary traits (Zhao et al. 2018). In tomato, a pan-genome that includes its wild relative species would provide a single coordinate system to anchor known nucleotide variation (SNP, InDels and CNV, for examples) with phenotypic data. The tomato genome reference was obtained from the Heinz1706 accession that experienced breeding during its history that leaded to fix or remove nucleotide variation. Thus, using a single reference genome is limiting the identification of novel genes from the available germplasm that are not present in this reference genome, especially for genes of agronomical interest. Rare CNV were already detected in the tomato genome demonstrating that structural variation exists in this species (Causse et al. 2013). In this context, it makes sense to re-think the idea of a ‘reference’ genome. The Pan-genome is also an opportunity to track chromosomal rearrangements between genotypes that may have occurred over micro (i.e. domestication) and macro (i.e. species divergence) timescales. While being computationally challenging, methodological approaches are available to construct, use and visualize pan-genome (The Computational Pan-Genomics Consortium et al. 2016).

### b. Modelling of demographic history and ecological niche

The genomes of contemporary crops contain considerable information about their history. Although, the general contour of tomato history has been defined with the increasing amount of available data (both SNP genotyping and sequencing) and sampling sizes, its resolution remains elusive. Statistical inference methods, inherited from human genomic and based on coalescent theory, have been developed to leverage information contained in these genome-wide sets of data

and have proven their power to refine parameters of the species history. More precisely, from observed footprints in DNA sequence variation, these methods aim at reconstructing the evolutionary history and providing precise estimates of selective and demographic events (i.e. population effective size growth or decline) that the species of interest experienced. Population genetic summary statistics (i.e. Watterson's Theta, Tajima's D) provide such data to test for demographic events. Numerous methods and models have been developed for demographic inferences (see Schraiber and Akey (2015) for a review) with the most popular ones being the principal component analysis (PCA), Structure (Falush et al. 2003) and Treemix software (Pickrell and Pritchard 2012) that are very powerful towards identifying population structure and mixture. In tomato, these methods have been largely applied and are the basis for further explorations of more complex demographic models that describe events like population divergence, migration and changes in demographic sizes. Towards this objective, methods based on site frequency spectrum (SFS) modelling have been applied in both the crop tomato (Lin et al. 2014) and its CWR species to unravel timings of population divergence for example (Beddows et al. 2017) but remain limited. Furthermore, until now, despite the large amount of genomic data, no haplotype-based method has been used to precisely measure coalescence between haplotype in a population to infer changes in its effective size, for example. The sequentially Markov coalescent (SMC) method and its extensions (PSMC (Li and Durbin 2011) and MSMC (Schiffels and Durbin 2014)), operating on full genome sequences, would be precious tool to precisely decipher this species history.

In parallel, past climate change may have contributed significantly to population dynamics and shaped patterns of nucleotide variation. Ecological niche modelling (ENM) building from current bioclimate variables are projected to paleoclimates to predict the variation in population geographical distribution over large time-scales. In tomato, the role of geography and ecology in species divergence has been investigated using a combination of climatic, geographic, and biological data from nine wild Andean tomato species to describe each species' ecological niche and to evaluate the likely ecological and geographical modes of speciation in this clade (Nakazato et al. 2008, 2010). Both studies mainly demonstrated that the nine studied species experienced an ecological adaptation that drove genetic and phenotypic divergence in association with one or more environmental variables, leading to specific ecological niches following a recent divergence. All these features turned these species into major source of biotic and abiotic stress-responsive genes and genetic mechanisms of adaptation to climate change. Those genetic resources can directly sustain

breeding efforts for elite germplasm that would grow under stressful or changing conditions without being detrimental to traits of economic interest such as yield or fruit quality.

### c. Adapting the tomato crop to climate change using genomic approaches

Food security may be threatened by a combination of events, such as increasing human population and needs, climate change and by the lack of sustainable development. Evolutionary adaptation has been proposed as a tool to understand how some species, such as the tomato, overcome environmental changes by the understanding of local adaptation mechanisms (Mousavi-Derazmahalleh et al. 2018). These changes act as selective pressures and are driven by climate change. However, the success of evolutionary adaptation depends on various factors, one of which being the extent of genetic variation available within the crop species. Many QTL studies have involved crop wild relatives, but just a few wild accessions were used (less than 10 *S. pimpinellifolium* and *S. habrochaites* and one or two of the other species, as reviewed by Grandillo and Cammareri (2016)). Thus, a large natural diversity, including important alleles for the crop, remains to be discovered and used to improve tomato adaptation. The genomic approaches provide a unique opportunity to identify genetic variation that can be employed for its own breeding efforts programs. The routinely use of genomic-based selection methods is a recent breakthrough facilitating the assessment of genetic variation and discovery of adaptive genes in this species. While additional information is needed, the current utility of selection tools indicates a robust ability to utilize existing variation in the tomato to address the challenges of climate uncertainty. Thus the objective is to properly use genomics to increase tomato yield, quality and stability of production through advanced breeding strategies, enhancing the resilience of this crop species to climate variability as proposed in Abberton et al. (2016).

### d. Implementing genome-wide based Genomic Selection

Genomic selection (GS) is a promising approach exploiting the density of molecular markers across genomes to offer advanced breeding designs (Goddard and Hayes 2007). More precisely, GS refers to selection decisions based on genomic breeding values (GEBV, Hayes et al. (2009)). This approach has the potential to be cost-effective (both in time and money) by reducing generation time or phenotyping effort through its prediction. While being successful in dairy cow breeding, its application in crops remains limited to major species such as maize (Crossa et al. 2013). This

methodological approach benefits from the availability of large genotyping or sequencing datasets, mainly obtained from GWA panels, to test its feasibility. The initial step, as described in Heffner et al. (2009), relies on performing a cross-validation (or model training cycle) step where the effect of parameters such as LD decay, size of the training population, density of markers on the correlation between the predicted phenotype and the measured phenotype are evaluated (the so-called '$r^2$' estimation). Using this knowledge, the most accurate prediction parameters and models can be determined. The cross-validation step offers the best framework to start with and run a first round of GEBV to select the best individual to reproduce.

In tomato, cross-validation studies have already been conducted providing an appreciation about the potential of GS in this species. The studied phenotypic traits were mainly related to fruit quality and showed a high predictability from a medium size GWA panel of nearly 160 individuals (accuracy up to 0.89 for fruit weight, (Duangjit et al. 2016)) but were variable according to the trait heritability: as expected, a low heritability trait was less predictable than high heritability trait. Additionally, the potential of GS was evaluated and showed that (1) reliable phenotype prediction models were constructed from simulation data leading to confident prediction for both yield and flavour, with for example, an $r^2$ of 0.807 for Solid Soluble Content (Yamamoto et al. 2016) and (2) quality traits improvement through GS can be reached for $F_1$ hybrid genotypes (Yamamoto et al. 2017). However, these studies also revealed that GS will be difficult to apply in a breeding context in tomato because of the number of traits to consider and the antagonism between fruit yield and quality traits (sugar content *vs* fruit size for example) combined with the high level of LD in modern varieties or the bottleneck of high-throughput phenotyping. But overall, tomato germplasm collections remain precious material that should be maintained, deeply characterized and enriched (notably with the addition of crop wild relative species) to support GS and GWA approaches.

### e. Opportunities from data sharing in the tomato scientific community

The past decade has been really fruitful in producing data such as genome sequences, transcriptomes and metabolomes. The type and quality of data may vary according to their generation technology (e.g. HiSeq vs PacBio, or RNAseq vs genome sequencing), and therefore it might be difficult to compare them within a same analysis. The real challenge is thus to develop databases that are user-friendly and help scientist handling the amount of data available. The tomato community with the creation of databases like Solgenomics Network (SGN - https://solgenomics.net/), tomatomics (Kudo et al. 2017), the Tomato Expression Atlas (Fernandez-

Pozo et al. 2017) and TomExpress (Zouine et al. 2017) has managed to acquire, collect and share most of the data available. It remains essential for researchers to make the best use of accumulated biological knowledge on tomato. In this context, the SGN database initiated a gathering of QTLs analyses but the discrepancy of alignment made it nearly unusable. Using methods developed in human in 2008 (Allen et al. 2008; Zeggini et al. 2008) and later applied in *A. thaliana* (Grimm et al. 2012), GWAS results were aggregated onto a cross-species platform to replicate results and share data. In tomato, many GWA studies have been conducted, especially on traits related to fruit quality, offering the opportunity to foster the genetic architecture of this trait through a GWA meta-analysis and consequently discover new candidate loci and reducing the proportion of uncovered heritability. This approach has notably been successfully applied in human (Tedja et al. 2018).

As we previously developed in this chapter, the higher nucleotide diversity from crop wild relative species will continue to supply breeding improvement. The data from crop and wild tomato species also represent an opportunity to expand scientific studies on plant biotic and abiotic stresses responses. Indeed, wild tomato species, being locally well adapted to all kind of extreme environments (from high altitude to arid areas), are a crucial resource for breeders to retrieve traits for future cultivars retaining high quality and performance despite environmental changes. Furthermore, high synteny revealed the common structure within families of plants such as for the *Solanaceae* (Wang et al. 2008; Rodriguez et al. 2009; Peters et al. 2012; Rinaldi et al. 2016) opening the possible diversity sources to the entire family. Therefore, useful discoveries in species like *S. melongena* or *S. tuberosum* could be translated to the tomato crop genome. This was recently demonstrated by the successful transfer of natural resistance from *Pisum sativum* to *A. thaliana* using new gene editing methods, such as CrispR-cas9 (Bastet et al. 2018), showing the promise of numerous future applications of this trans-specific process. At the opposite, another potentially successful approach to sustain the development of high-yielding crops was recently proposed and could be applied in tomato. This approach, called the ‘rewilding’, consists in furnishing crops that carry lost properties that the ancestors once possessed to tolerate variable environmental conditions (Palmgren et al. 2015).

## Further Major Readings We Recommend:

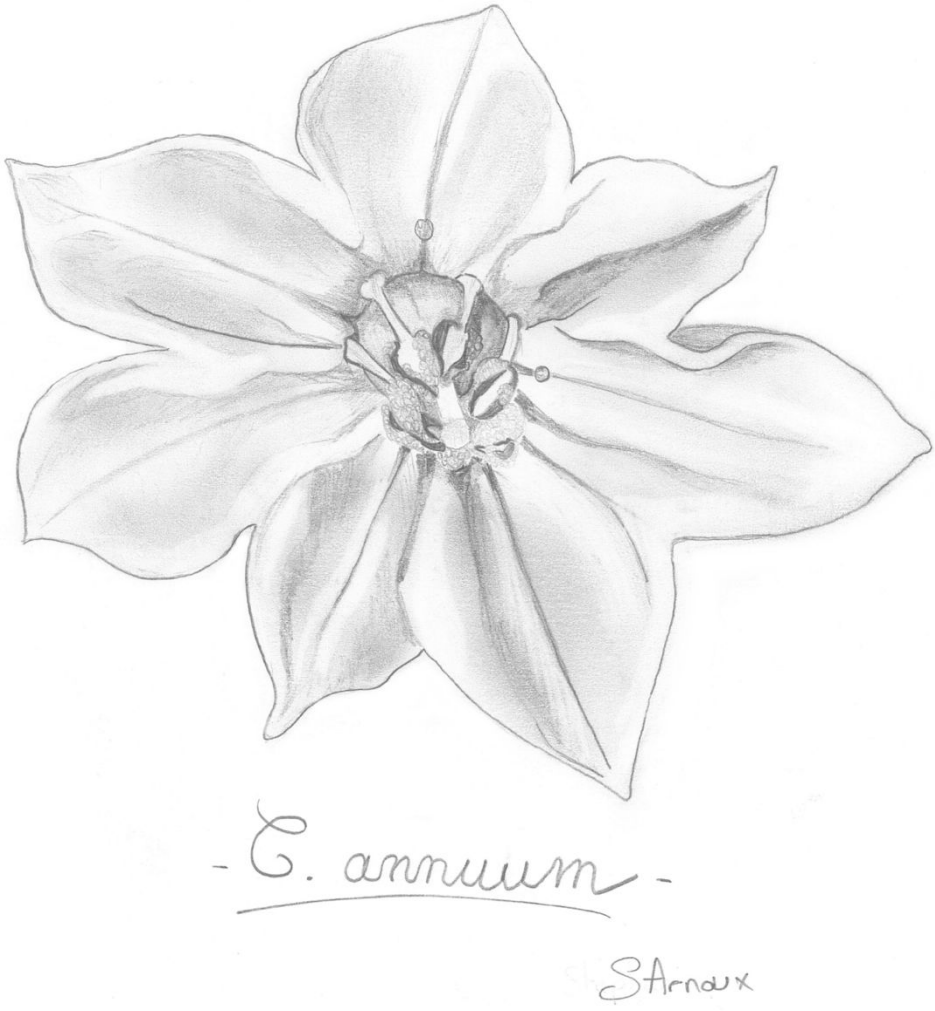
- Bauchet and Causse 2012 : 'Genetic Diversity in Tomato (Solanum lycopersicum) and Its Wild Relatives' (Bauchet and Causse 2012)

- Haak et al. 2014 : 'Merging Ecology and Genomics to Dissect Diversity in Wild Tomatoes and Their Relatives' (Haak et al. 2014)

- Peralta, I; Spooner, D; Knapp, S 2008 Taxonomy of wild tomatoes and their relatives (Solanum sect. Lycopersicoides, sect. Juglandifolia, sect. Lycopersicon; Solanaceae) Amer. Society of Plant Taxonomists, 2008 (Peralta et al. 2008)

# CHAPTER 2

## Demographic inferences reveal a convergence of domestication in *Solanaceae*



In the second chapter, we aimed to decipher the most likely domestication scenario for the three crop and wild population pairs. We performed a comparative analysis of several demographic models of increasing complexity to limit biases induced by making strong assumptions (Gaut et al. 2018). Comparing the crop and wild populations enabled us to evaluate the extent of biological changes due to domestication. This knowledge is crucial to improve future breeding efforts, and we bring valuable estimation of the impact of human selection on the crop effective population size and gene flow with their wild relative (Zeder 2015). Inferring the demographic scenarios of these three species is an unprecedented opportunity to further characterize each domestication event duration, and therefore improve the inference of the demographic history that were hypothesized through indirect means (human and cultivation history of the areas, ancient written records).

Results in brief:

The comparative study of the demographic inferences modeling the domestication of the three species, revealed the convergence of the domestication processes in the Solanaceae family

- Detection of **artificial selection** footprints in the Solanaceae **genomes**
- Presence of a **bottleneck** corroborating with the domestication **stage of cultivation**, in the three species
- Estimation of the **divergence time** between the crop and their wild relatives
    - **Eggplant** domestication: 5,938-3,087 BCE
    - **Pepper** domestication: 6,760-3,514 BCE
    - **Tomato** domestication: 7,901-4,107 BCE

Conclusion and perspectives:

- By knowing the past behavior of our crops facing domestication events, we improve modern breeding efforts to sustain future crop breeding and their innate barriers to human control conditions
- Possible applications for producing *de-novo* domestication events

# Demographic inferences reveal a convergence of domestication in *Solanaceae*

Manuscript submitted to Molecular Ecology

Stéphanie Arnoux[1], Christelle Fraïsse[2], Christopher Sauvage[1]*

[1], INRA UR1052 GAFL, Centre de Recherche INRA PACA, Domaine Saint Maurice, 67 Allée des Chênes, CS60094, 84140 Avignon Cedex 9, France

[2], Institute of Science and Technology Austria, Am Campus 1, Klosterneuburg 3400, Austria.

* Corresponding author

## Chapter 2

### *Demographic inferences reveal a convergence of domestication in Solanaceae. - 93 -*


**Abstract** **- 97 -**

**1. Introduction** **- 99 -**

**2. Materials and Methods** **- 100 -**
Plant Materials and RNA sequencing ..... - 100 -
Quality control, reads alignment and variant calling ..... - 101 -
PCA and unfolded allele frequency spectrum..... - 102 -
Demographic inferences ..... - 102 -

**3. Results** **- 103 -**
Biological material and high-throughput sequencing ..... - 103 -
Genetic structure and Allele frequency spectrum..... - 104 -
Demographic inferences and best scenario choice..... - 108 -
Parameter estimates and bootstrap of each species' best model ..... - 108 -

**4. Discussion** **- 110 -**
Domestication footprints on Solanaceae genomes: impact of the artificial selection ..... - 110 -
A convergent bottleneck revealed by the comparative analysis of three Solanaceae species..... - 111 -
Timing of the different stages in the domestication process ..... - 112 -
Conclusion ..... - 114 -

**Acknowledgements** **- 114 -**

**Data Accessibility Statement** **- 115 -**

**Author Contributions** **- 115 -**

**Supplementary Figures** **- 116 -**

**Supplementary Tables** **- 117 -**

## Abstract

Domestication is a human-induced selection process that imprints the genomes of domesticated populations over a short time scale. Deciphering whether these changes are convergent between independent domestication histories needs to be ascertained. Reconstructing historical gene flow and effective population size changes is therefore of fundamental interest to understand how demography and human selection jointly shaped genomic divergence during domestication. Here we used an extended modeling framework based on demographic divergence models that capture temporal variation in effective population size and migration rate to explore the multiple facets of domestication-with-gene-flow. We investigate the domestication history of three pairs of species of Solanaceae (eggplant, pepper and tomato) characterized by distinct domestication history, including geographic isolation from the wild progenitor for pepper and tomato and sympatry for eggplant. RNAseq derived SNPs were used to document the extent of genetic differentiation in each species pairs, and ten different models were fitted and compared based on the unfolded joint allele frequency spectrum of each pair. We found evidence of bottleneck in the three species. Our results also suggest that the timings of domestication of these three species are supported by the few historical records available. This study thus provides a new retrospective insight into the historical demographic process that shapes Solanaceae through domestication and further promotes the hierarchical fitting of increasingly complex demographic models.

**Keywords:** demographic inference, site frequency spectrum, domestication, population genomics, Solanaceae.

## 1. Introduction

Domestication involves a few thousand years of human selection that represents a great opportunity to understand evolution (Darwin 1868; Diamond 2002). The domestication process has been described as following four stages (Meyer and Purugganan 2013). The process starts as a management of wild populations favoring particular phenotypes (stage 1), followed by their cultivation (stage 2) often resulting in a genetic bottleneck due to the separation of the cultivated crops and their wild progenitor. The crop plants are then dispersed worldwide and need to adapt to new local conditions, via introgressions from crop wild relative or new mutation fixations (stage 3), and finally there is a deliberate breeding effort that includes crosses of modern cultivars (stage 4). The main interest for modern breeding is to understand the crop responses to artificial selection, i.e. the domestication syndrome (Hammer 1984; Vigne 2011). Especially the stage 2 of domestication is of particular concern as it often results in a bottleneck in plants, especially in annuals (Miller and Gross 2011). Indeed, a few wild plants with specific phenotypes were selected for traits such as flowering time (Blackman et al. 2011), plant architecture (Clark et al. 2004) and fruit size (Frary et al. 2000). This selection for favorable alleles imprints the whole genome, as shown in maize (Hufford et al. 2013), rice (Caicedo et al. 2007; Nabholz et al. 2014) and tomato (Koenig et al. 2013; Sauvage et al. 2017). A better knowledge and dating of this stage 2 is directly linked to human history, as it started with the settlement of human populations and the beginning of crop cultivation (Zeder 2015).

With the combined use of genomic resources and demographic inferences between crop and wild progenitors, it becomes possible to detect and characterize these stages, by changes in the effective population size and gene flow rate, and estimate their duration (Gaut et al. 2018). While most demographic studies reconstruct the domestication events in a single species (Eyre-Walker et al. 1998; Wright 2005; Zhu et al. 2007; Sabeti et al. 2007), the question of the convergence between the domestication processes among different crop species received little attention so far. Here, we took advantage of the common selected phenotypes in three domesticated species of the Solanaceae family to evaluate the extent of convergence between independent domestication events. This taxonomic family is composed of several species of major scientific and economical interest, such as potato, tomato or tobacco, for which large genetic and genomic resources are available.

We selected three species (eggplant, pepper and tomato) with different geographical origins and for which reference genome sequences are available. The crop eggplant, *Solanum melongena L.*, was domesticated in Asia (Meyer et al. 2012b) and it is only recently that *S. insanum* was proposed

as its wild progenitor (Aubriot et al. 2016; Ranil et al. 2016). Both species remain in sympatry within Asia, but the range of eggplant production and consumption expanded worldwide (Davidar et al. 2015). The crop pepper, *Capsicum annuum L.*, is bred and consumed worldwide and is native to tropical Mesoamerica. It was domesticated in Mexico (Perry et al. 2007; Ibiza et al. 2012) before being introduced in Europe (Andrews 1993). The supposed common wild progenitor, C. *annuum var. glabriusculum*, shows high discrepancy in its phylogeny and remains not well defined (Hill et al. 2013; Nicolaï et al. 2013). The cultivated tomato, *Solanum lycopersicum* L., was domesticated from the wild progenitor *S. pimpinellifolium* in Peru before experiencing two bottlenecks: first moving to Mesoamerica (Jose Blanca et al. 2012) and then with few cultivars introduced to Europe from Mexico (Atherton and Harris 1986; Blanca et al. 2012). These domestication events and further genetic improvement led to specific genomic footprints in tomato (Lin et al. 2014).

As previous studies, we aimed to decipher the most probable domestication scenario for the three crop and wild population pairs (Nabholz et al. 2014; Qi et al. 2017). We compared several demographic models of increasing complexity to limit biases induced by making simplifying assumptions while avoiding overfitting (Gaut et al. 2018). Comparing the crop and wild populations enabled us to evaluate the extent of biological changes due to domestication. Indeed, it is crucial to estimate the impact of human selection on the crop effective population size and gene flow with their wild relative to bring insights into future breeding improvement efforts (Zeder 2015). Furthermore, it is an unprecedented opportunity to further characterize the duration of each domestication phase, and therefore improve the inference of the demographic history that were hypothesized through indirect means (human and cultivation history of the areas, ancient written records).

## 2. Materials and Methods

### Plant Materials and RNA sequencing

To complete the RNAseq data available in Pease *et al.* 2016 (3 accessions of tomato) and Sauvage *et al.* 2017 (8 accessions of tomato), we sampled crop and wild accessions for three species within the Solanaceae family (eggplant, pepper and tomato). All accession details are given in the Table S1. For each species, accessions were selected according to the literature to maximize nucleotide diversity within the crop population and their wild relatives. Seeds were collected from the INRA seed bank from the Genetic Resources Center

(https://www6.paca.inra.fr/gafl_eng/Vegetables-GRC/). For eggplant we used seven crop accessions (*S. melongena*) and 10 wild accessions (*S. insanum*) (Aubriot et al. 2016); for pepper we used 12 crop accessions (*C. annuum*) and four accessions of the supposed common ancestor *C. annuum var. glabriusculum* (Hernández-Verdugo et al. 2001; Qin et al. 2014). For the tomato we used nine crop accessions (*S. lycopersicum* – eight previously used in Sauvage *et al.* 2017 and one used in Pease et al. 2016) and nine wild relative accessions from the close relative species *S. pimpinellifolium* (Blanca et al. 2012, 2015). For the ancestral states, we used 2 outgroup accessions in eggplant (*S. campylacanthum*), 4 outgroup accessions in pepper (*C. microcarpum*, C. *baccatum* and *C*. *chacoense*) and 3 outgroup accessions in tomato (*S. lycopersicoides* and S. *sitiens* both retrieved from Pease *et al.* 2016).

Three replicates of each accession were grown in greenhouses under normal conditions during spring and summer 2012, in Avignon, France. The biological samples were pooled, and composed of 15, 20 and 65% of flower, fruit and leaf tissues, respectively. Briefly, these different tissues were chosen to catch the broader representation of gene expression levels for the entire plant. Entire flowers and young leaves were sampled while fruits were harvested only at ripe stage (40 days post anthesis). All tissues were flash frozen in liquid nitrogen prior to storage at -80°C and subsequent RNA extraction using the Spectrum Plant Total RNA from SIGMA-ALDRICH (ref. STN50), following manufacturer's recommendations. RNA obtained was quantified and its quality was checked using a bioanalyser 2100. RNAseq libraries were prepared and individually tagged (using 6 bp tags) at INRA SupAgro (Montpellier, France) using the TrueSeq kit and sequencing was performed by the GetPlage Platform (INRA, Toulouse), using the HISEQ2500 protocol (150 bp stranded and paired-ends reads were produced). The transcriptomic analyses are described in Arnoux *et al.* 2018.

### Quality control, reads alignment and variant calling

We performed sequencing data quality control using FastQC and trimmed the adapters from the sequences using Trimmomatic (Bolger et al. 2014b). The sequences of all accessions were aligned to the respective reference crop transcriptome, for eggplant: *S. melongena* (The Eggplant Genome Consortium 2017; Lanteri et al. 2014), for pepper: *C. annuum* (Qin et al. 2014) and for tomato: *S. lycopersicum* (The Tomato Consortium 2012). We used a python pipeline (*cf data availability* section - GitHub repository) to perform the mapping on the respective reference set of coding sequences using BWA-MEM (Li 2013). GATK was used to call the variants (HaplotypeCaller), perform base quality score recalibration, indel realignment and duplicate removal according to the GATK Best Practices

recommendations (DePristo et al. 2011; Van der Auwera et al. 2013). VCFtools (Danecek et al. 2011) was applied to filter the output variant calling file (vcf) and retained sites showing a minimal coverage per individual of 20x and minimal mean coverage over the total set of accessions mapped of 10x. Following the SNP calling, we used the approach implemented in reads2snp (Gayral et al. 2013) to make a clean cut off of paralogous sites selecting for FIS under 0.5. As genetic linkage is increased during domestication due to a higher selfing rate (Charlesworth and Wright 2001), Plink (Purcell et al. 2007; Chang et al. 2015)was used to remove linkage disequilibrium (LD) by pruning the linked SNPs (with a $r^2 > 0.4$). In tomato, we excluded the chromosome nine of any subsequent analysis as it was almost entirely introgressed from a wild relative accession of *S. peruvianum* to bring resistance to Tomato mosaic virus (*Tm2.2* locus), and therefore would have biased our analyses (Ohmori et al. 1995; Koenig et al. 2013).

### PCA and unfolded allele frequency spectrum

A principal component analysis (PCA) was performed using the filtered pedigree file (ped) on the SNP genotype data from the wild and crop populations (Fig. 1). Then, the ancestral status of each SNP was determined from the consensus of the outgroup accessions. The 4P software (Benazzo et al. 2015) was subsequently applied to produce a unfolded joint allele frequency spectrum (jAFS) that combines the unfolded AFS of the crop and wild populations of each species. As wild and crop seeds were kept under greenhouse' conditions for tens generations, the level of inbreeding coefficient (FIS) was inflated. The presence of highly inbred individuals in our populations does not fit the requirement of the demographic inferences. Thus, we circumvented this issue by selecting randomly one of the two alleles of each individual, joining two random individuals together, and creating a diploid-like population with all the genetic details but no inbreeding. All the scripts and bioinformatic procedures are detailed in the depository on GitHub (https://github.com/starnoux/arnoux_et_al_2019).

### Demographic inferences

We estimated the joint demography of the wild and crop populations for each Solanaceae species using the maximum likelihood approach implemented in a modified ∂a∂i version 1.6.3 (Gutenkunst et al. 2009). In total, we defined 10 demographic models to test: (i) the timing of gene flow during divergence between wild and crop populations (absence of gene flow: SI, continuous and asymmetric gene flow: IM, two periods (early and late) of continuous and asymmetric gene flow:

IM2); (ii) the sequence of population size changes (constant population size: C, gradual growth or decline: E, bottleneck: B); (iii) and their number (one, two or three periods). All models began with the split of the ancestral population in two daughter populations, and then were followed by a sequence of demographic events in the absence or presence of gene flow. A summary of the models is given in the Figure S1, and the scripts used to define models in ðaði are provided in the section 5.dadi_inference of the GitHub repository (see above).

For each model, we ran 50 independent runs from randomized starting parameter values, and for each run we performed two rounds of optimization. A global optimization ("simulated annealing" method) from the randomized starting values was followed by a local optimization ("BFGS" method) starting from the optimized values in the previous step. To assess the relative support of models with different number of parameters, we used the Akaike Information Criterion (AIC), calculated as 2**k* - 2**logL*, where *k* is the number of parameters of the model, and *logL* its maximum log likelihood value across the 50 independent runs. Model comparisons are described in Table S2 and represented in the Figure S1.

We set the bounds of the prior for each parameter according to historical records in the three species: population size changes, times, migration rates and proportions are detailed in the Table S3. The inferred parameter values were scaled by the effective population size of the ancestral population calculated as $N_A$ = *theta* / (4 * *mu* * *L*), where *theta* was inferred by ðaði; *mu* is the mutation rate per nucleotide per generation estimated to be between $1x10^{-08}$ (higher bound) and $5.20x10^{-09}$ (lower bound) as suggested in wild tomato and higher plants (Moniz De Sá and Drouin 1996; Dvornyk et al. 2002; Roselius et al. 2005; Lynch 2010) and *L* is the length of sequenced DNA, i.e. the filtered transcriptome length, equals to 19,468,437 bp in eggplant, 18,401,318 bp in pepper; and 20,160,440 bp in tomato. Solanaceae species being annual plants, we consider a generation time of 1 year in our duration estimations. The converted parameter estimates of the best model for each Solanaceae species, and its 95% confidence intervals obtained with the Godambe methods (Coffman et al. 2016) from 1000 conventional bootstraps over SNPs, are given in Table S4.

## 3. Results

### Biological material and high-throughput sequencing

We generated RNAseq data for crops, wild relatives and outgroups of three Solanaceae species: (i) eggplant: 7 crop accessions (S. *melongena*), 6 wild accessions (S. *insanum*) and 2 outgroup

accessions (S. *campylacanthum*); (ii) pepper: 10 crop accessions (C. *annuum*), 4 wild accessions (C. *annuum* var. *glabriusculum*) and 4 outgroup accessions (C. *chacoense* and C. *baccatum*); (iii) 9 crop accessions (S. *lycopersicum*), 9 wild accessions (S. *pimpinellifolium*) and 3 outgroup accessions (*S. sitiens* and *S. lycopersicoides*). No significant differences in the mapping quality was observed, with a percentage of read mapped ranging from 74% to 81% in eggplant, from 68% to 75% in pepper and from 76% to 85% in tomato (details of the mapping statistics are provided in Table S5). Reads were assigned to 96.8% of the genes in eggplant (33,209 over 34,396), 97.9% of the genes in pepper (34,610 over 35,336) and 95.8% of the genes in tomato (34,297 over 35,768). We obtained 727,629 SNPs in eggplant, 1,061,975 SNPs in pepper and 2,912,381 SNPs in tomato. After filtering for paralogs, and LD pruning, we based our analyses on 16,955 SNPs in eggplant, 41,508 SNPs in pepper and 33,535 SNPs in tomato.

### Genetic structure and Allele frequency spectrum

From the filtered SNPs, we assessed the genetic relationships between crop and wild individuals in each species by performing a PCA based on the genotype data (Fig. 1). For the three species, the crop accessions represented by colored diamonds were clustered together, and separated from the wild accessions in grey circles. However, the crop and wild populations still have an overall low level of genetic differentiation, as shown by the modest fraction of genetic variance explained by the first PCA axes. Moreover, the genetic distance between the wild accessions is consistently higher than for the crop accessions, whatever the pair of species considered (as shown in the Chapter 3). Especially, the wild pepper population differs from the crop on the first three axes of variance, and therefore it has a quite strong genetic structure.

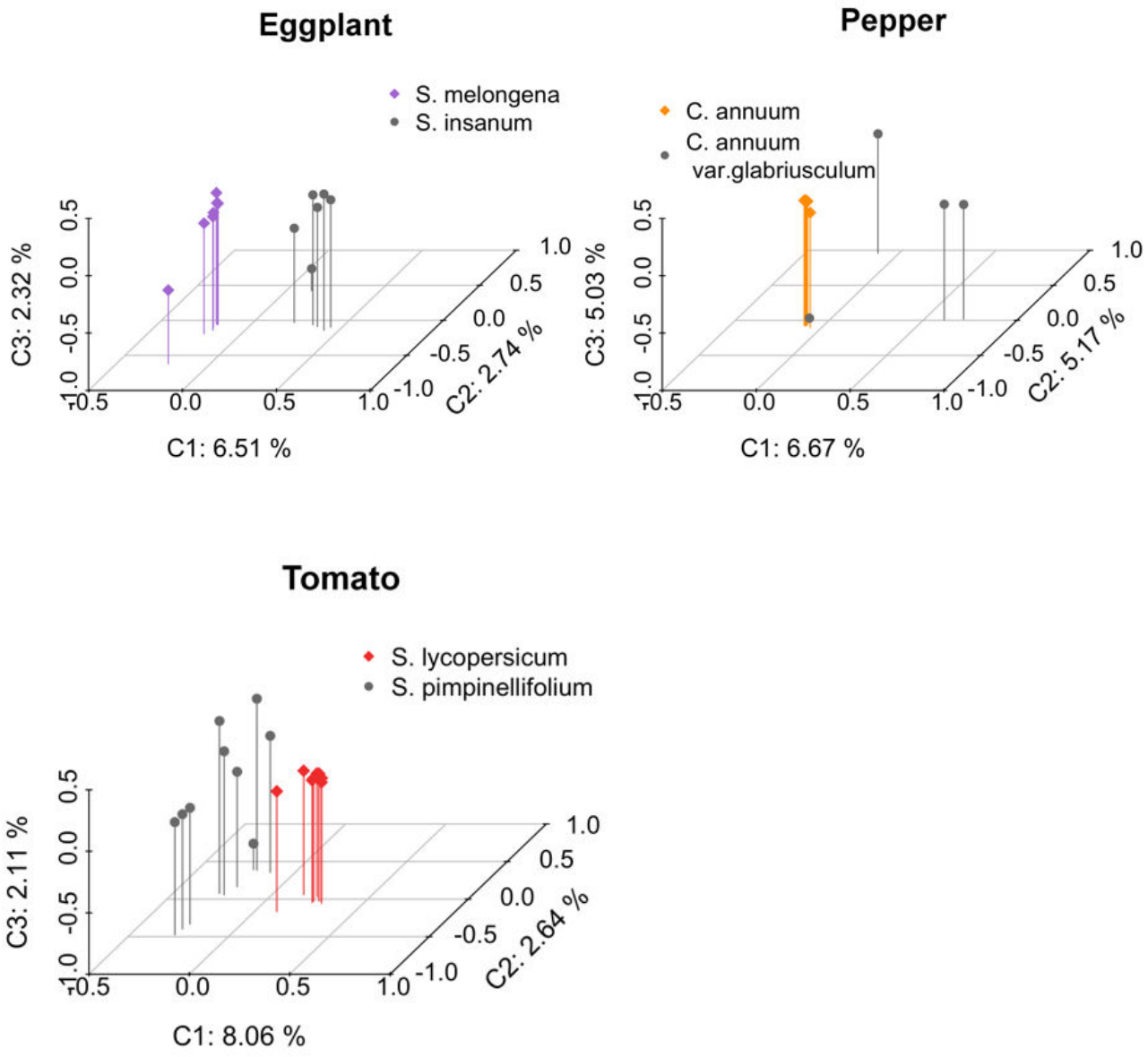



**Figure 1.** Graphical representation of the principal component analysis based on genetic covariance among all accessions (crop: colored diamonds; wild: grey circle) for each species. The first three principal components are represented with the fraction of variance explained.

The unidimensional allele frequency spectrum of each species between the crop and wild populations was produced from this same set of SNPs. And we observed a high level of inbreeding in all the populations except for the wild tomato. This issue was solved by rearranging our data to conform with the random mating assumption made in our inferences (see Material & Methods). By using the consensus of outgroup sequences, we polarized 12,977 SNPs in eggplant (76.53% of the filtered set), 38,296 SNPs in pepper (92.26% of the filtered set) and 26,135 SNPs in tomato (77.93% of the filtered set). These oriented SNPs were used to produce a joint allele frequency spectrum (jAFS) of the crop and wild populations for each species (Fig. 2*b*).

Using the ∂a∂i approach, we explicitly modeled domestication in the three Solanaceae species (Fig. 2). We first applied a model choice procedure between various demographic scenarios to test whether (i) the crop population was connected to the wild population during its domestication history, (ii) the crop population had experienced a strict bottleneck (strong reduction in the effective population size at a time point) or a gradual reduction of its population size, (iii) the domestication event was concomitant with the reduction in population size. Results are detailed in the Table S6 which reports the posterior estimates of the two best supported scenarios for each species (see Table S2 for full details across all models). Some models better supported were corrupted and/or had posteriors stumbling over boundaries due to overfitting, these were discarded. For all three species we observed unambiguous support in favor of bi-directional gene flow during the whole divergence history between the crop and wild populations (except during early divergence in tomato), and a reduction of effective population size in the crop population (either a bottleneck (tomato and pepper) or a gradual decrease (eggplant)). This period was followed by an effective population size expansion in the crop populations of eggplant and tomato, but not in the pepper (Fig. 2*a*).

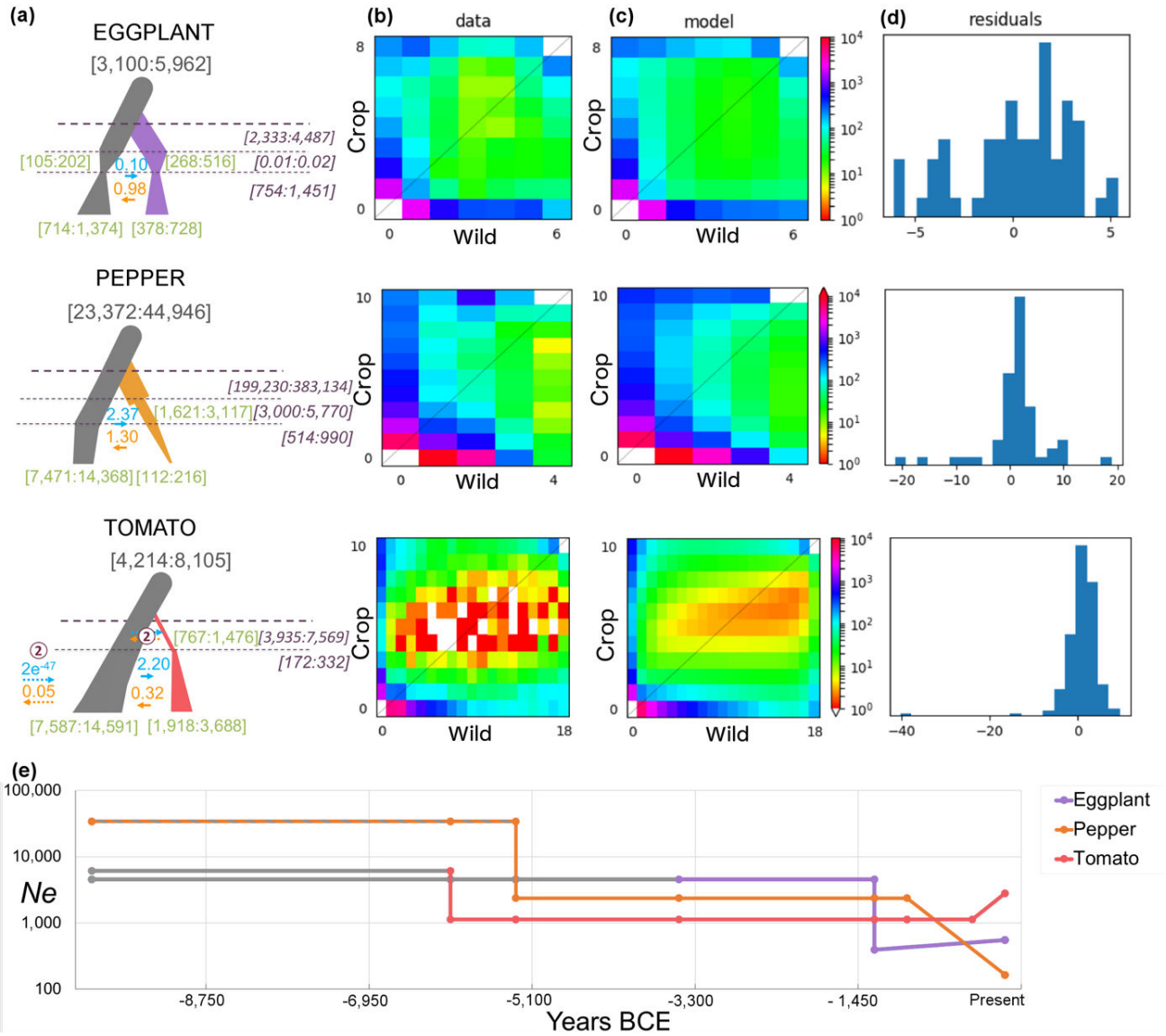



**Figure 2.** Historical demography of the three Solanaceae species. (a) Best model for each species, with the parameter values estimated using the established mutation rate range (min=5.20x10$^{-09}$ and max=1x10$^{-08}$). Black: effective size of the ancestral population; green: effective size of the crop and wild populations; purple: timing of the demographic events (in years); blue: migration rate from wild to crop (migrants per generation); orange: migration rate from crop to wild (migrants per generation). (b) Observed joint allele frequency spectrum (jAFS) for wild (*x* axis) and crop (*y* axis) populations for each species. For each jAFS the color scale represents the number of SNPs falling in each bin defined by a unique combination of the number of derived alleles observed in crop and wild populations. (c) Predicted jAFS of the best model for each species. (d) Histogram of the residuals between the best model and the data for each species. (e) Graphical representation of the changes in the effective size (*y* axis, log scale Ne) over time (*x* axis), with parameters averaged between the upper and lower estimates using the range of mutation rate, for the crop population of the three species.

### Demographic inferences and best scenario choice

For the three species, the best model was supported with a much higher AIC than the second best-fitting model (Fig. 3), though posterior parameter estimates were comparable between each other's (Table S6). The best models were: (i) in eggplant, IM_C_E_E (Isolation with Migration, one period of constant population size and two successive events of gradual change in population sizes, Fig. S1) with an AIC of 845 (log-likelihood = -412.9164); (ii) in pepper, IM_C_BcCw_E (Isolation with Migration, one period of constant population size followed by one period of bottleneck in the crop and a successive event of gradual change in population sizes, Fig S1) with an AIC of 2361 (log-likelihood = -1171.982); (iii) and in tomato, IM2_C_BcCw_E (the same model as pepper, except that migration was negligible in the early divergence, Fig S1) with an AIC of 4499 (log-likelihood = -2239.9959). The Table S7 provides the biological conversion of the parameters from the best and second-best supported model estimates, and they are detailed in the Figure S2.

### Parameter estimates and bootstrap of each species' best model

We inferred the parameter estimates and their confidence interval (all bootstrap estimates are detailed in the Table S4) under the best scenario for each species . The effective population size of the ancestral population, from which the crop and wild populations diverged, was approximately 3,100-5,962 in eggplant, 23,372-44,946 in pepper and 4,214-8,105 in tomato; the two estimates stand for the lower/upper bound of the probable range of mutation rates (see Material & Methods). The wild populations in the three species experienced different demographic scenarios. The wild eggplant experienced a strong reduction in the effective population size followed by an expansion (*Ne* from 3,100-5,962 to 105-202 and then to 714-1,374). The wild pepper only experienced a single reduction in the effective population size (*Ne* down to 7,471-14,368) and the tomato wild population followed an expansion (*Ne* up to 7,587-14,591).

Concerning the crop populations, the three species experienced a strict bottleneck or a strong reduction in the effective population size over species-specific duration. The decrease in the effective population size in eggplant was almost instantaneous and was followed by an expansion during 754-1,451 years (*Ne* from 3,100-5,962 to 268-516 and then to 378-728). In pepper, the split between the crop and the wild populations was really old (199,230-383,134 years ago) and the first bottleneck occurred 3,514-6,760 years ago and lasted for 3,000-5,770 years until a more recent decrease in the effective population size occurred in the past 514-990 BCE (*Ne* from 23,372-44,946 to 1,621-3,117

and then to 112-216). Similarly, tomato first experienced a severe bottleneck 4,108-7,901 BCE followed by a constant period over 3,935-7,569 years, and then an expansion over 172-332 years (*Ne* from 4,214-8,105 to 767-1,476 and then to 1,918-3,688). The crop and wild populations split time was estimated to be 3,088-5,939 years in eggplant, 202,745-189,895 years in pepper and 4,108-7,901 years in tomato.

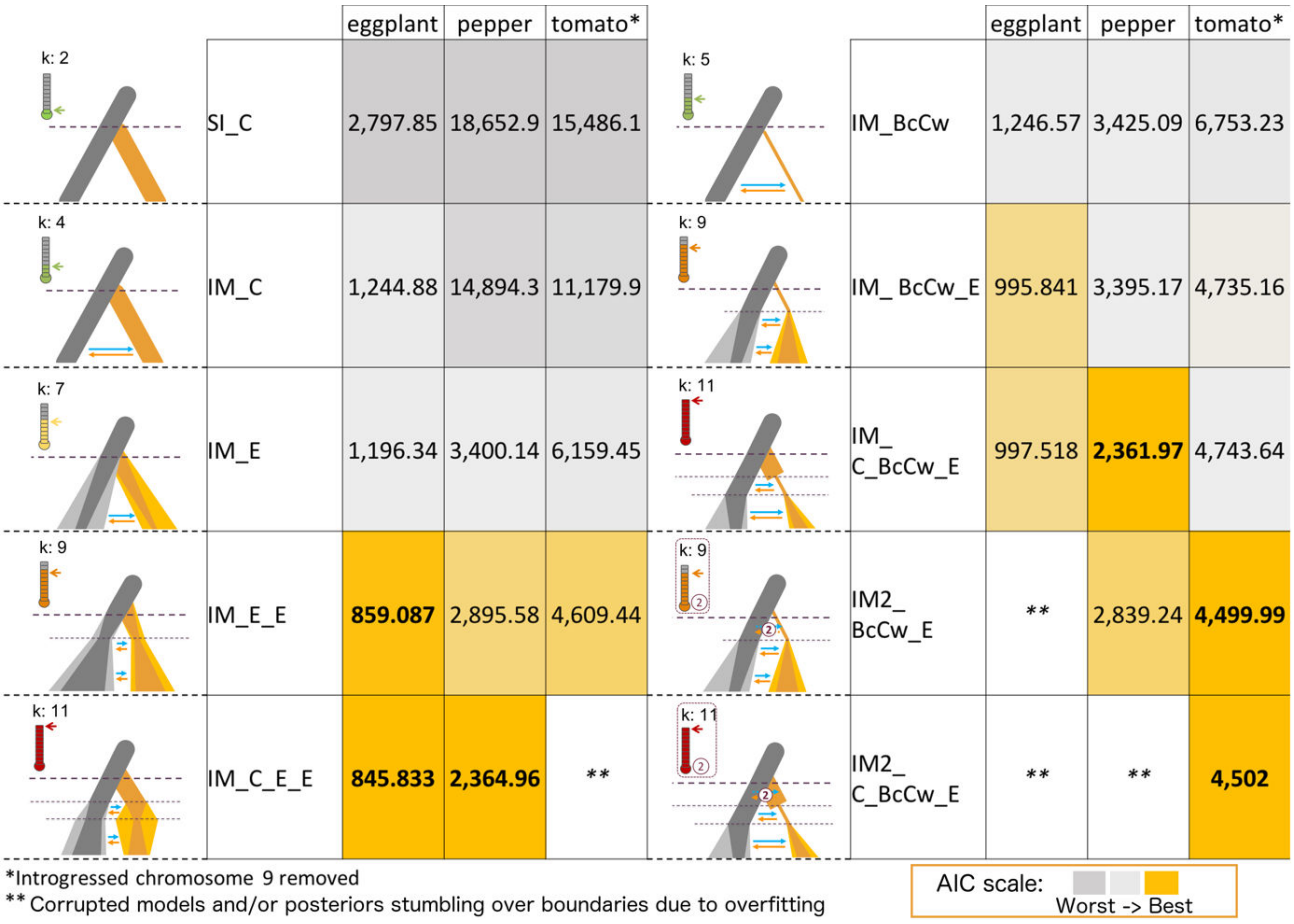



**Figure 3.** Heat-map of the AIC values for the 10 demographic models showing the best inference for each species. The number of model parameters (*k*) is scaled from green for simple models to red for complex models. Warmer colors indicate better models.

Gene flow between the crop and wild populations was ongoing during the whole divergence history, and asymmetric: eggplant experienced a smaller level of gene flow from wild to crop (0.10

vs. 0.98), while the contrary was true for the pepper (2.37 vs 1.30) and the tomato (negligible migration during the early divergence and 2.20 vs 0.32 during the demographic expansion), in agreement with their degree of geographic isolation.

Together our comparative analyses confirmed that the strong genetic diversity erosion observed in the crop populations was due to a reduction in their effective population size during domestication in all three species, though details of the demographic events differ between them. Moreover, our results support the idea that domestication started during a comparable period in the three Solanaceae species. This convergence in the domestication process is notable as the different species were domesticated independently in different geographic regions.

## 4. Discussion

Our demographic analyses of more than 10,000 oriented SNPs of crop and wild populations, clarified the domestication process of three Solanaceae species. We found strong evidence for convergence in the demographic impact of domestication in the three species, and yet, divergent scenarios of migration between crop and wild were inferred in agreement with their degree of geographic isolation (allopatry vs. sympatry). Most importantly, our study is the forerunner of comparative demographic inferences between species of the same clade (see also Jouganous *et al.* 2017 for a recent method that compares multiple population demographic history). The hierarchical fitting of increasingly complex models allowed us to test different domestication scenarios and refine the duration of the different domestication phases in eggplant, pepper and tomato.

### Domestication footprints on Solanaceae genomes: impact of the artificial selection

Domestication in plants reduced the nucleotide diversity genome-wide through artificial selection (Caicedo et al. 2007; Hufford et al. 2013; Koenig et al. 2013; Lin et al. 2014; Nabholz et al. 2014; Sauvage et al. 2017). This selection was paired with environmental and demographic changes, that imprinted the genome. The best demographic model for each of the three species in our study revealed a reduction in crop effective size, from 7% to 18% of the ancestral effective population size. This drastic and rapid, almost instant, reduction referred as bottleneck is found in the domestication events of the three species. The modern breeding following the dispersion of domesticates had different effects in each species. In pepper the decrease in crop effective size seems to occur until recently, as reported in sorghum where modern breeding deteriorates the genetic diversity throughout the history of cultivation and this until present days (Smith et al. 2018). In tomato we

removed the chromosome nine that was totally introgressed in some accessions (Young and Tanksley 1989; Lin et al. 2014). In both eggplant and tomato an increase in crop effective size arose most probably due to introgressions and modern crosses with wild individuals (Atherton and Harris 1986; Ibiza et al. 2010). This last point is important to better understand the evolution of domestication, since bottlenecked populations undergoing demographic expansion are more likely to carry and even fix slightly deleterious alleles (Luikart 1998; Excoffier et al. 2013; Peischl et al. 2013; Lohmueller 2014). Often referred as the cost of domestication, it was first described as the increase in non-synonymous substitutions in domesticated compared to wild lineages of rice (Lu et al. 2006; Glémin and Bataillon 2009; Moyers et al. 2018).

In the case of eggplant, we detected migration from the crop to the wild population mostly due to their high level of outcrossing and them being in sympatry (Meyer et al. 2012b). The human-mediated selfing of crop plants to retain phenotype of interest in garden could enhance their isolation, and explain this asymmetric gene flow (e.g., Brandvain et al. 2014). Domestication is increasing inbreeding in crops, and this recurrent artificial selection would act as a barrier to natural introgressions from the wild.

During domestication of both tomato and pepper, the environmental conditions for crop landraces has been totally shifted. The landraces moved to man-controlled environment, which were non-native and characterized by totally different environmental conditions (Loaiza-Figueroa et al. 1989; Blanca et al. 2012). The fitness associated with their developmental traits changed consequently, and modern breeding is mostly responsible for the asymmetric gene flow that they both experienced. This gene flow from wild to crop seemed to be the detectable footprint of human-mediated introgressions (Atherton and Harris 1986; Ibiza et al. 2010; Chitwood et al. 2013).

### A convergent bottleneck revealed by the comparative analysis of three Solanaceae species

In the past, inferences on domestication often relied on a predetermined demographic model with simplifying assumptions about duration and effective size changes, as pointed out in Gaut et al. (2018). Here, using the jAFS as a summary statistic of genome-wide differentiation, we compared 10 scenarios of increasing complexity that model the demographic process of domestication in our three species. To increase the power of the method and generate asymmetric distributions of derived variants around the jAFS diagonal, each jAFS was oriented with wild related species as outgroup. Concerning the temporal variation in effective size, we did not force the reduction to be instantaneous (though bottlenecks were also implemented). In eggplant, the models implementing

bottlenecks were consistently corrupted, while the best model included a drastic reduction in effective size over a short duration (i.e. similar to a bottleneck). This confirm the hypothesis of a "second stage domestication" (i.e. cultivation) (Meyer and Purugganan 2013) occurring in a rapid time frame (Ladizinsky 1987; Zohary 1989). In annual plants, and especially in our three species, this stage follows a bottleneck model that leads to the fixation of domestication alleles (Doebley 1989a; Miller and Gross 2011).

So far, few efforts were made to implement the possibility of migration between the crop and the wild populations, except recent studies in cereal crops (Caicedo et al. 2007; Molina et al. 2011; Beissinger et al. 2016). Following these attempts, we implemented models including an asymmetric migration rate constant through the entire process of domestication and modern breeding. The isolation with migration (IM) scenario provided best fit to the observed data for the three species pairs. To this model we add the option of a second migration rate, modelling the specific exchanges between the crop and wild populations, during the first demographic event. This relaxation of assumptions reveals that crop tomato was totally separated from its wild compartment at first, and then, migration from wild to crop population occurred in the last 172 to 332 years, which corroborates perfectly with modern efforts for breeding improvement based on wild introgressions that intensified with modern breeding since the beginning of the 20$^{th}$ century (Pimentel et al. 1997; Brozynska et al. 2015). Another feature detected is the high genetic divergence between wild and crop pepper populations. Unexpectedly, the wild species is way more divergent than a wild progenitor and it reveals a lack in our botanical knowledge of the wild progenitor of *C. annuum* from our samples. A clarified and strongly supported phylogeny of the *C. annuum* remains needed.

### Timing of the different stages in the domestication process

In addition to providing a better knowledge of domestication our results also imply a deeper look at human history. In this demographic study we focused on the domestication stages starting from the cultivation to the modern breeding. Surprisingly the duration of domestication remains unclear for most plants until now (Gaut et al. 2018). Accounting for temporal variation in migration rate and effective size allowed us to better estimate the duration of the different stages of domestication, with a quite high certainty, by assuming a range of possible mutation rate. A critical question was to know if the bottleneck, and therefore the cultivation stage of domestication, occurred directly after the split between crop and wild populations, or if we could detect a protracted

period of management before the cultivation as it was shown in African rice and grape using similar demographic inference methods (Li and Durbin 2011; Schiffels and Durbin 2014; Terhorst et al. 2017).

In our case, from the splitting time between the crop and the wild population, tomato shows a direct bottleneck and no protracted period. In pepper, having wild species further apart than the wild progenitor removes the possibility of testing. Only the eggplant has a clear protracted period preceding the bottleneck, which suggests that human management favored particular phenotypes before the separation of the domesticated and the wild genotypes. Though, we were not able to detect a second domestication event with our analysis as previously proposed (Meyer et al. 2012b). When other plants, such as maize, have archeo-botanical records (Piperno et al. 2009; Ranere et al. 2009), in Solanaceae the seed storage doesn't allow a conservation of sufficient quality and only few papers related such records in *Capsicum spp.* without ascertaining the species (Duncan et al. 2009; Kraft et al. 2014). Therefore, our parameter estimations are of first importance. In eggplant, the protracted stage started around 5,938-3,087 BCE, which corroborate with old writing records already describing crop phenotypes of eggplant 3,200-2,600 BCE (Suśruta and Bhishagratna 1907). Then, the cultivation period that follows the deep decrease in effective size 1,451-754 BCE would support the cultivation and export of eggplant towards Japan and middle-east in the 8$^{th}$ century BC and the strong gene flow with wild populations (Daunay and Laterrot 2007). Pepper was domesticated probably in eastern/central Mexico or in the Yucatan Peninsula region (Aguilar-Melendez et al. 2009), among the lima beans (Martínez-Castillo et al. 2007) and upland cotton (Brubaker and Wendel 1994). At the opposite of other plants such as eggplant and tomato, pepper populations seem to have strongly reduced in size over the recent hundred years 990-514 BCE. This might be the result of a strong world-wide extensive selection for specific phenotypic traits such as pungency level (Pickersgill and Heiser 1977), and likely to have occurred after the discovery of the new world (Eshbaugh 1993). We estimate the first cultivation event and bottleneck to occur around 6,760-3,514 BCE. This timing corroborates the estimate of the age of the Mayan and Oto-Manguean languages (about 6,000 BCE), that already used names to designate pepper ca. 80 (Kaufman 1994). It also agrees with pepper seed remnants found in caves at Tehuacan, Mexico that dated about 7,000 BCE but without certainty about the species (Yamaguchi 1983). The tomato domestication center is located in Peru (Blanca et al. 2012), even though it was imported into Mexico further increasing the bottleneck effect due to cultivation. We estimated this bottleneck to date 7,901-4,107 BCE which would fit the first area cultivation records for maize in Peru 6,775-6,504 BCE (Grobman et al. 2012) and in Mexico 7,300 BCE (Pohl et al. 2007). Unfortunately, we would need further sampling specific from Peru to evaluate if the

geographic origin of the first domestication event occurred in Peru or in Mexico (Jose Blanca et al. 2012). This type of approach has been successfully applied in the African rice using spatially explicit coalescent simulation and whole-genome sequences to shade light on the geographical origin of the crop , thus deciphering the domestication center (Cubry et al. 2018).

## Conclusion

Overall, our study provides insights into the convergence of the domestication processes in the Solanaceae family. While the geographic and demographic dimensions might differ among the different annual species, we observed convergent bottleneck events produced by the domestication stage of cultivation. Our results also point out the relevance of such comparative demographic inferences to decipher the estimates of population sizes, migration rates and timings at the intraspecific level between the crop and wild populations. Together, these inferences bring new details about the timing of domestication and therefore about human history. It confirms the importance of understanding how plant species respond to human manipulation. By knowing the past behavior of our crops facing domestication events, we improve modern breeding efforts to sustain future crop breeding and their innate barriers to human control conditions (Zeder 2015).

## Acknowledgements

This work was supported by the EU Marie Curie Career Integration grant (FP7-PEOPLE-2011-CIG grant agreement PCIG10-GA-2011-304164) attributed to CS. SA was supported by a PhD fellowship from the French Région PACA and INRA, in partnership with Gautier Semences. CF was supported by an IST fellow (Marie Sklodowska-Curie Co-Funding European program). Authors thanks Mathilde Causse and Beatriz Vicoso for their team leading. Thanks to the Italian Eggplant Genome Consortium, which includes the DISAFA, Plant Genetics and Breeding (University of Torino), the Biotechnology Department (University of Verona), the CREA-ORL in Montanaso Lombardo (LO) and the ENEA in Rome for providing access to the eggplant genome reference. Thanks to CRB-lég (https://www6.paca.inra.fr/gafl_eng/Vegetables-GRC) for managing and providing the genetic resources, to Marie-Christine Daunay and Alain Palloix (INRA UR1052) for assistance in choosing the biological materiel used, to Muriel Latreille and Sylvain Santoni from the UMR AGAP (INRA Montpellier, France) for their help with RNAseq library preparation, to Jean-Paul Bouchet and Jacques Lagnel (INRA UR1052) for his Bioinformatics assistance.

## Data Accessibility Statement

- Raw sequence data generated for the three species samples used in these analyses are hosted at the EMBL European Nucleotide Archive under project number PRJEB26324.
- All the procedures (e.g. mapping pipeline, model inferences scripts) associated packages and software versions used for the study are detailed in the GitHub repository, [https://github.com/starnoux/arnoux_et_al_2019]
- VCF are available upon request.

## Author Contributions

CS planned and designed the study. SA produced the figures. SA and CF performed the analyses and wrote the manuscript. CS supervised the study and the writing.

## Supplementary Figures

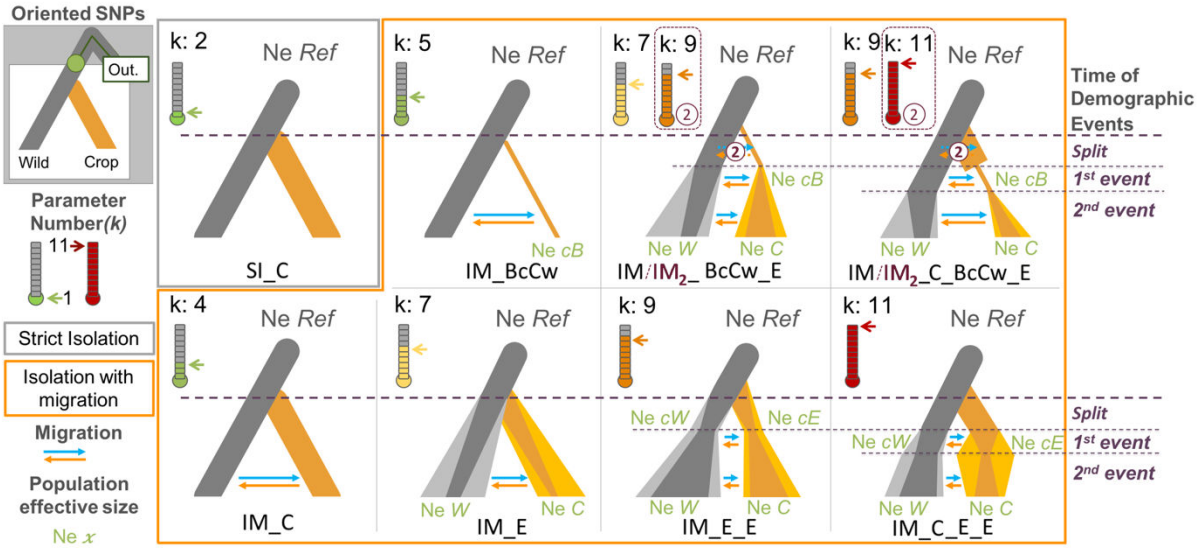



**Figure S1.** Graphical representation of the 10 demographic models implemented in this study. The grey box shows the simple model of strict isolation (SI), while the yellow box represents various models of Isolation with Migration (IM). Briefly, *Ne* correspond to the effective population size (*Ne W*: wild, *Ne C*: crop, *Ne cB*: crop after bottleneck, *Ne cW*: wild after growth/decline, *Ne cE*: crop after growth/decline), migration is shown by orange (from crop to wild) and blue (opposite direction) arrows, and the number of model parameters (*k*) is scaled from green for simple models to red for complex models.

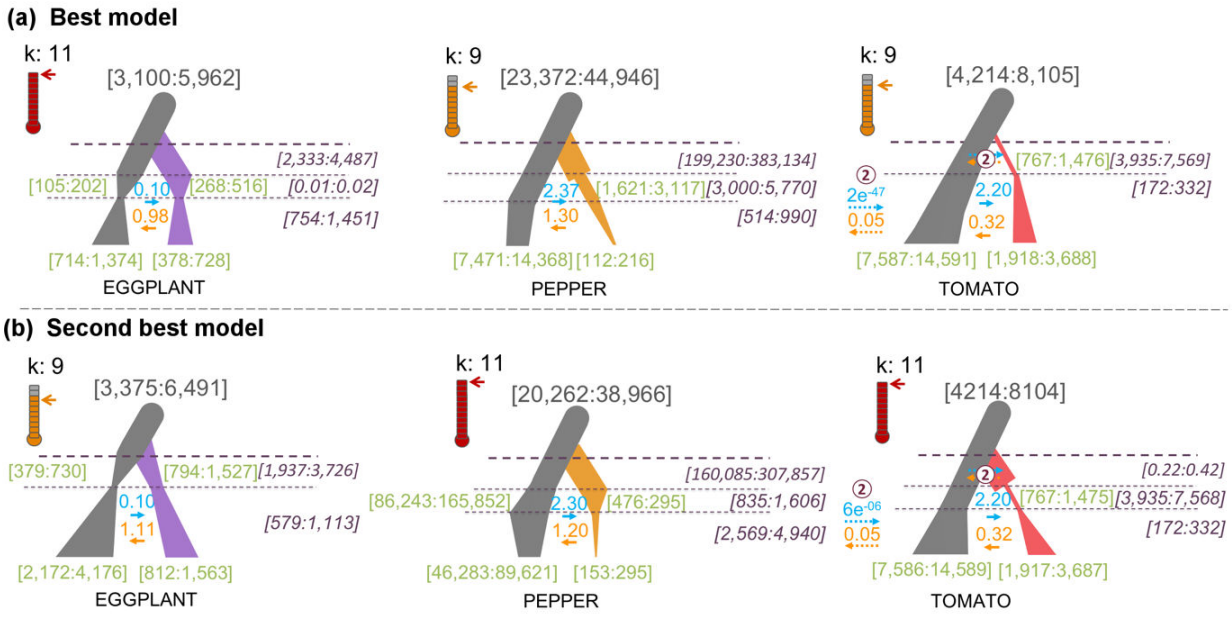



**Figure S2.** Representation of the two fittest models for each species. Parameter estimates are indicated using the established mutation rate range (min=5.20x10-09 and max=1x10-08). The number of model parameters (k) is scaled from green for simple models to red for complex models. Other details match Figure 2.

## Supplementary Tables

The supplementary tables are available in the appendix 2 and are composed of the following tables:

**Table S1:** Detailed data about the studied accessions. The species and the location of origins are listed in separated tables for the three accessions: a. the eggplant accessions, b. the pepper accessions and c. the tomato accessions. The accessions in white background are the crop species, the ones in grey background are the wild species and the ones in green background are the outgroups.

**Table S2:** The best of the 50 runs is selected according to the log likelihood and the detailed posterior parameters are listed, for each species, for the 10 models. The models in green are the best models, and the pink values are the posteriors stumbling over boundaries due to overfitting. *k* is the number of parameters of the model; *n* is the number of finished inferences of the 50 independent runs; *AIC* is the Akaike Information Criterion (calculated as 2**k* - 2**logL*); *Ne* correspond to the effective population size relative to the *Nref* (*Ne W*: wild, *Ne C*: crop, *Ne cB*: crop after bottleneck, *Ne cW*: wild after growth/decline, *Ne cE*: crop after growth/decline); m correspond to the migration rate (*mCW*: migration rate from wild to crop, *mWC*: migration rate from crop to wild); *T* represents the times in generations relative to the *Nref* (*Ts*: duration of the first epoch from the split to next demographic event, *Tb*: duration of the second epoch, *Te*: duration of the third epoch); *Theta* is related to the *Nref*, the length of the sequences used to obtain the jAFS and to the mutation rate.

**Table S3:** Detailed boundaries and prior probabilities are listed for each parameter for each species. All other details match Table S2.

**Table S4:** Detailed bootstraps results (x1,000) of the Godambe method, on the two best models of each species. For each parameter, the best estimate and its standard deviation obtained by bootstrap is provided. The best model is indicated in the first row, the second-best in the second row. All other details match Table S2.

**Table S5:** Mapping summary statistics on mapped, properly paired and singletons reads of all the studied accessions aligned to their reference genome, in the three species. The accessions in white background are the crop species, the ones in grey background are the wild species and the ones in green background are the outgroups.

**Table S6:** Detailed posterior parameters of the two best models for each species. The best model is indicated in the first row, the second-best in the second row. All other details match Table S2.

**Table S7:** Biological conversion of the estimated parameters for the two best models for each species. All ðaði output parameters are given in the white backgrounded table. All parameter conversions or estimates are given in a range of two possible mutation rate (min=$5.20 \times 10^{-09}$ and max=$1 \times 10^{-08}$) in the yellow backgrounded table. For each parameter, the best estimate and its standard deviation obtained by bootstrap is provided. The estimated effective size is given as population size and not as ratio and the duration are estimated in generation (in annual plants: 1 generation = 1 year). All other details match Table S2.

# CHAPTER 3

## Domestication footprints reveal a convergence of both nucleotide diversity and gene expression in cultivated Solanaceae

In the third chapter, the hypothesis is a convergent modification of gene diversity and gene expression during domestication. Comparing crop and wild relative accessions enabled to estimate gene expression differences and detect genomic selection footprints. Annotations of the targeted genes (selected and differentially expressed) identified the biological processes altered during domestication. The hypothesis relies on the orthologs shared within the trio of species and their modification. We hypothesize that mechanisms of regulation and adaptation that have been triggered by domestication of crop species are convergent. Therefore, for the three independent domestication process the expectation is to highlight parallel changes induced in crops compare to their wild relatives.

Results in brief:

The study of orthologs highlighted a convergence of the molecular changes for the three species,

- at the genetic level:
    - **Relaxation of selection:** transcription initiation, translational initiation and tolerance to abiotic stresses
    - **Direction selection**: plant growth and fruit development
- at the gene expression level:
    - **Down-regulation**: regulation of abiotic responses and drought tolerance
    - **Up-regulation:** plant-growth, cell expansion, leaf growth, fruit development and ripening

Conclusion and perspectives:

- Pathway impacted by domestication are global and therefore impact more polygenic pathways than local genes
- Deeping the study could lead to detect specific genes co-expressed involved in domestication
- Possible applications for retrieving adaptive traits such as drought tolerance from the wild populations
- Real concern to conserve wild populations as diversity sources

# Domestication footprints reveal a convergence of both nucleotide diversity and gene expression in cultivated Solanaceae

Manuscript to be re-submitted,

The review from Genome, Biology and Evolution are provided in the appendix 4

Stéphanie Arnoux, Renaud Duboscq, Mathilde Causse and Christopher Sauvage*

INRA UR1052 GAFL, Centre de Recherche INRA PACA, Domaine Saint Maurice, 67 Allée des Chênes, CS60094, 84140 Avignon Cedex 9, France

* Corresponding author

# Chapter 3

***Domestication footprints reveal a convergence of both nucleotide diversity and gene expression in cultivated Solanaceae - 119 -***

**Abstract - 123 -**

**1. Introduction - 125 -**

**2. Material and Methods - 126 -**

Plant Materials ..... - 126 -
Alignment pipeline ..... - 127 -
Nucleotide diversity ..... - 128 -
Differential expression analyses ..... - 128 -
Annotations and orthology analyses across the three species ..... - 129 -
Statistical analyses ..... - 129 -
Data availability ..... - 130 -

**3. Results - 130 -**

Biological material ..... - 130 -
Identification of genetic diversity shifts ..... - 132 -
Identification of differentially expressed genes ..... - 133 -
Annotations and enrichment analyses ..... - 134 -
Statistical analyses ..... - 137 -

**4. Discussion - 139 -**

**Acknowledgements - 143 -**

**Author contributions - 144 -**

**Supplementary Figures - 145 -**

**Supplementary Tables - 147 -**

## Abstract

The conscious and unconscious selection induced during the domestication and modern breeding stages of crop history led to considerable phenotypic and genetic changes. Studies focused on major effect genes associated to domestication by studying polymorphisms or gene expression. In the present study we explore the convergence of both processes in three cultivated *Solanaceae*. To identify domestication convergence, we compare the genetic diversity and gene expression levels between crop and wild accessions in a trio of species. We analyze the transcriptomes of 47 genotypes, including wild and landraces of tomato (cult. *Solanum lycopersicum*; wild *S. peruvianum*), eggplant (cult. *S. melongena*; wild *S. insanum*) and pepper (*Capsicum annuum*; wild *C. chilense* and *C. frutescens*). Across the three species, the magnitude of differential expression levels revealed a convergent rewiring of genes during the domestication that are in congruence with the ones targeted by selection. In addition, the expressed genes log fold change variation was significantly correlated with the nucleotide diversity variation, in the three species. While our transcriptomic analyses confirmed the changes in expression of numerous domestication related genes, the novelty of our study is the highlight of the convergence of domestication footprints acting on both nucleotide diversity and gene expression.

Key words: Evolutionary transcriptomic, modern breeding, domestication, *Solanum lycopersicum*, *Solanum melongena*, *Capsicum annuum*.

## 1. Introduction

Domestication of plants and animals appeared with the settlement of human populations and the beginning of farming (Zeder 2015). A few wild plants were selected according to their phenotype such as flowering time (Blackman et al. 2011), plant architecture (Clark et al. 2004) and fruit size (Frary 2000) commonly called domestication syndrome (Hammer 1984; Vigne 2011). This selective process is a rich model to study evolution and adaptation. At the molecular level, the selection for favorable alleles induced a genetic bottleneck that imprinted the whole genome, as shown in maize (Hufford et al. 2013), rice (Caicedo et al. 2007; Nabholz et al. 2014) and tomato (Koenig et al. 2013; Sauvage et al. 2017). This selection was paired with a relaxation of natural selection on traits that lost importance in the crops (Innan and Kim 2004). The changes in nucleotide diversity due to selection, came with a rewiring of gene expression levels, as observed in maize (Wright 2005), tomato (Koenig et al. 2013; Sauvage et al. 2017) and common bean (Bellucci et al. 2014). Studies demonstrated the magnitude of the induced perturbation expanding to complete pathways shutdown (Itkin et al. 2013) to remove anti-nutritional compounds.

Two types of footprints can be tracked, shifts in nucleotide diversity and gene expression level changes. The question of their correlation and convergence through domestication events received little attention. Almost no studies properly tested this convergent hypothesis, especially in related species. To test for convergent trans-specific signatures of selection on nucleotide diversity and gene expression levels, we took advantage of the parallel history of domestication in the Solanaceae family. This family is composed of several species of major scientific and economical interest, such as potato, tomato or tobacco. We chose three species, eggplant, pepper and tomato for which highly colinear and syntenic genomes are available (Wang et al. 2008). They all experienced similar phenotypic selection during independent domestication and modern breeding stages and have different geographical origins. Eggplant is originating from Africa where wild species are still present (Knapp et al. 2013; Meyer et al. 2015). Wild species of eggplant moved to Asia where *Solanum melongena* L. was domesticated (Meyer et al. 2012b). Though, it is only recently that *S. insanum* was proposed to be the closest common ancestor or wild progenitor of the crop eggplant (Aubriot et al. 2016; Ranil et al. 2016). Both species remain in sympatry within Asia, but the range of eggplant production and consumption expanded worldwide (Davidar et al. 2015). The pepper crop, *Capsicum annuum* L. is bred and consumed worldwide and is native to tropical Mesoamerica. It was

domesticated in Mexico (Perry et al. 2007; Ibiza et al. 2012) before being introduced in Europe (Andrews 1993). The complex of cultivated species *C. chinense* and *C. frutescens* was used as outgroup to the *C. annuum* crop in Hill et al. (2013) because the supposed common wild progenitor (*C. annuum* var. *glabriusculum*) shows high discrepancy in phylogeny (Hill et al. 2013; Nicolaï et al. 2013). These two species are commonly considered as admixing and sharing the same locations in the lower Andean region (Pickersgill 1971; Walsh and Hoot 2001; Guzmán et al. 2005). Both were used in *C. annuum* breeding as source of resistance to diseases (Polston et al. 2006; Ibiza et al. 2010) and pests (Fery and Thies 1997). The cultivated tomato, *Solanum lycopersicum* L. was domesticated in Peru before experiencing two bottlenecks, first moving from Peru to Mesoamerica (Blanca et al. 2012) and then through the introduction of a few cultivars from Mexico to Europe (Atherton and Harris 1986; Blanca et al. 2015). Even if tomato has been domesticated from the wild progenitor *S. pimpinellifolium*, the wild relative species from the peruvianum Clade (Pease et al. 2016) have been used and remain source of genetic diversity especially for disease resistances (Ohmori et al. 1995; Lin et al. 2014).

Using these *Solanaceae* species, we investigate the convergence induced by domestication and modern breeding at the molecular level. Comparing crop and wild relative accessions enables to estimate gene expression differences, detect genomic selection footprints and test for their correlation. Annotations of the targeted genes identify the biological processes altered during domestication. Indeed, it is crucial to decipher the induced changes in crops and the remaining sources of wild relatives' diversity. Comparing three species is an unprecedented opportunity to decipher convergent mechanisms of regulation and adaptation that have been triggered by domestication of crop species.

## 2. Material and Methods

### Plant Materials

To conduct our comparative genomics approach, we sampled crop and wild accessions (hereafter called population pairs) for three species within the Solanaceae family, eggplant, pepper and tomato. All accessions were selected according to the literature and description in the seed bank of the genetic resources to get a range representing the nucleotide diversity within the crop and the wild populations. For eggplant we used seven crop accessions (*S. melongena*) and 11 wild accessions

(*S. insanum*) (Aubriot et al. 2016), for pepper we used 11 crop accessions (*C. annuum*) and four accessions of close relative species that were source of diversity for improvement of *C. annuum* and are both well clustered phylogenetically further apart from *C. annuum* (*C. frutescens* and *C.* chinense) (Carrizo García et al. 2016) and for the tomato we used eight crop accessions (*S. lycopersicum* – previously used in Sauvage et al. 2017) and six wild relative accessions from 3 species of the *Peruvianum* clade (*S. peruvianum*. *S. huaylasense* and *S. corneliomulleri)* as defined in Pease et al.) (see Table S1 for the detailed description of the sequencing data).

The plants were grown under glasshouse's conditions with three replicates per accession. The biological samples were composed of sampled tissues pooled according to respectively a 15, 20 and 65% proportion of flower, fruit and leaf tissues to get the broadest representation of gene expression levels for the entire plant. Fruit samples were harvested at the ripe stage (40 DPA), while entire young leaves were sampled. All tissues were flash frozen in liquid nitrogen before storage at -80°C and subsequent RNA extraction using the Spectrum Plant Total RNA from SIGMA-ALDRICH (ref. STN50), following manufacturer's recommendations. RNAseq libraries were prepared and individually tagged (using 6 bp tags) at INRA SupAgro (Montpellier, France) using the TrueSeq kit and sequencing was performed by the GetPlage Platform (INRA, Toulouse), using the HISEQ2500 protocol (150 bp stranded and paired-ends reads).

### Alignment pipeline

We performed sequencing data quality control using FastQC and trimmed the adapters from the sequences using Trimmomatic (Bolger et al. 2014b). The sequences of each species, crop and wild populations, were aligned to the respective crop reference genome, for eggplant: *S. melongena* (Lanteri et al. 2014; The Eggplant Genome Consortium 2017), for pepper: *C. annuum* (Qin et al. 2014) and for tomato: *S. lycopersicum* (The Tomato Genome Consortium 2012). We used a python language pipeline to perform the mapping on the respective reference set of CDS using BWA-MEM (Li 2013). The Haplotype caller from GATK (HaplotypeCaller) called the variants according to GATK Best Practices recommendations (DePristo et al. 2011; Van der Auwera et al. 2013). The VCFtools (Danecek et al. 2011) filtered the output variant calling file to retain sites showing a minimal coverage per individual over the total set of accessions mapped of 20x. We used the approach implemented in reads2snp (Gayral et al. 2013) to make a clean cut off of paralogous sites.

## Nucleotide diversity

Once the paralogous sites were filtered, we produced principal component analysis (PCA) on the SNP genotype data with the R package 'SNPRelate' (Zheng et al. 2012; R Core Team 2016). Using DNAsp (Rozas et al. 2017), we estimated the total nucleotide diversity (π) per gene within each populations of the 3 species which is a relative measure of the degree of polymorphism within a population that can be used to detect balancing selection and hard sweeps (Hohenlohe et al. 2011). The π estimates of each population were plotted genome-wide, and the values were smoothed over 50 genes with the 'rollMean' function in R.

We removed outlier nucleotide diversity values considered as remaining paralogous sites. This extreme nucleotide diversity limits were set at 5% of the tail of the distribution. Then, we plotted for each species the $\pi_{CROP}$ against the $\pi_{WILD}$ of all the filtered genes. We examined if genes experienced severe shifts of nucleotide diversity between the crop and the wild population, hereafter denominated shifted genes. To detect these shifted genes, we defined two thresholds, (i) one filtering the highest π values using the 0.95 quantile (in eggplant: high π values thresh.$_{CROP}$: $1.2\times10^{-3}$, thresh.$_{WILD}$: $1.8\times10^{-3}$ ; in pepper: thresh.$_{CROP}$: $2.1\times10^{-3}$, thresh.$_{WILD}$: $6.4\times10^{-3}$ ; in tomato: thresh.$_{CROP}$: $9.8\times10^{-4}$, thresh.$_{WILD}$: $6.2\times10^{-3}$) and the second (ii) filtering the lowest π values using a $1\times10^{-3}$ of the maximum value in the crop population as it is the lowest non-null value (low π values in eggplant thresh.: $1.8\times10^{-4}$ ; in pepper thresh.: $1.1\times 10^{-4}$ ; in tomato thresh.: $7.8\times10^{-5}$). Once these thresholds were defined, the π shifted groups of genes were split into: (i) the group A with genes highly diverse in the crop population, showing a relaxation of selection or diversifying selection in the crop, and (ii) the group B with genes with almost no nucleotide diversity in crop but with a high nucleotide diversity in wild population, indicating directional selection in the crop group.

## Differential expression analyses

The estimation of the raw read counts (RC) per gene was obtained using the Samtools *idxstats* option (Li 2011) and the table of RC per accession and per gene (Table S2a., b and c for each species) was produced using a homemade R script (c.f. paragraph on Data availability). RC were normalized with a regularized log transformation for each accession to get gene expression levels for the subsequent analyses. A PCA was performed on normalized gene expression data transformed by the variance stabilization to show global patterns of gene expression between groups of individuals. To identify significantly differentially expressed genes (DEG) between the crop and the wild population

for each species, we used the statistical framework implemented in the R package *DEseq2* (Love et al. 2014) with a false discovery rate of 1%. Thus, the up- and down-regulated genes (defined as the ratio of gene expression levels of the crop over the wild population) were detected and Log Fold Changes (LFC) were assigned to each DEG for each pair of species.

### Annotations and orthology analyses across the three species

To reduce the inherent bias due to the heterogenous functional annotation quality, we reannotated transcriptomes following an identical approach. The Interproscan annotation system (Jones et al. 2014) was followed to retrieve the gene ontologies (GO) from the Pfam library (Finn et al. 2016) of the three reference transcriptomes, allowing a consistent comparison of gene ontologies between species. For each species, we used the gene ontologies to detect biological processes (BP) over-represented in the set of DEG and shifted genes (crop-diverse and wild-diverse groups treated separately) compared to the total expressed genes. We used the Wallenius non-central hypergeometric distribution implemented in the R/Bioconductor package 'goseq' (Young et al. 2010) and used the '*eval.go*' function as described in Sauvage et al. (2017). *P-value* thresholds of 0.1 was applied to test for the enrichment in shifted genes and 0.05 for DEG genes. In each species, for both sets of tested genes, the gene space used was the entire set of expressed genes for which a GO term was assigned. All the parameters are detailed in the GitHub {cf. Scripts and detailed parameters}.

We identified the 1:1 (across a pair of species) and the 1:1:1 (across the three species) orthologous genes with the software 'proteinortho' (Lechner et al. 2011). Similarly, from these sets of 1:1 and 1:1:1 orthologs, we performed GO enrichment analyses using test sets of genes composed of the DEG and the shifted genes (groups crop- and wild-diverse) that overlapped the set of orthologs. By this means, we tested whether genes orthologous in each of the three species and differentially expressed (between crop and wild) or shifted were enriched in any GO.

### Statistical analyses

For each expressed gene, using a GLM procedure, we tested the correlation between the pressure of selection and the gene expression changes induced by domestication in the three species. We used generalized linear models (GLMs) as they can handle non-normal responses by using the variance of each measurement, as an attribute of the response, as a linear function of covariates. With GLMs we could consider the chromosome effects as a regression factor and detect the actual

relationships between $\Delta\pi$ and LFC. Thus, we used $\Delta\pi(\pi_{WILD} - \pi_{CROP})$ (a proxy for changes in selective pressure, only with non-zero values) and correlated these with LFC estimates. The regression estimates for each chromosome were compared between each other, using the first chromosome as reference, in order to detect chromosomes that would have experienced a different nucleotide diversity change. We tested the correlation between (i) the changes in gene expression levels and the changes in nucleotide diversity and (ii) the reciprocal model. Following are the linear models we used:

(i) $\Delta\pi(\pi_{WILD} - \pi_{CROP}) \sim LFC + Chromosome$

(ii) $LFC \sim \Delta\pi + Chromosome$

We used the 'Gaussianize' function from the R package LambertW (Goerg 2011) to transform the $\Delta\pi$ data and LFC that were expected to deviate from normal distributions. In our four models, the GLM tested the chromosome effect and then took it into account if significant ($p<0.05$). We ran the GLM models under R (version 3.3.3) using the package lme4 (Bates et al. 2015). Additionally, for each species, we used a pairwise Fisher's exact test ($p<0.05$) to detect if the proportion of DEG or shifted genes per chromosome was significantly different from the proportion of DEG (or shifted genes) in the genome.

### Data availability

All the procedures and scripts (packages and software versions) used for the study are in the GitHub repository (https://github.com/starnoux/arnoux_et_al_2018). Raw sequences data used in these analyses are hosted at the European Nucleotide Archive under project number PRJEB26324.

## 3. Results

### Biological material

We generated RNAseq data for crops and wild relatives of the three pairs of species: seven and 11 accessions of crop (*S. melongena*) and wild (*S. insanum*) eggplant, 11 and four accessions of crop (*C. annuum*) and wild (*C. frutescens* and *C. chinense*) pepper and eight and six accessions of crop (*S. lycopersicum*) and wild (Peruvianum clade: *S. peruvianum*, *S. huaylasense* and *S. corneliomulleri*) tomato. No significant differences in the mapping rate was noticed with a percentage of read

mapping ranging from 74% to 81% in eggplant, from 68% to 75% in pepper and from 76% to 85% in tomato (Detailed mapping statistics are provided in the Table S3). Reads were assigned to 96.8% of the genes in eggplant, 97.9% of the genes in pepper and 95.8% of the genes in tomato. The table S4 is providing both the number of genes where at least one raw read was mapped and the number of CDS we used for subsequent analyses.

The variant caller detected 727,629 SNPs in eggplant, 1,061,975 in pepper and 2,912,381 in tomato. After quality and paralog controls, we based our analyses on 112,773 SNPs in eggplant, 213,683 in pepper and 950,036 in tomato. These sets of SNPs were located in 17,545, 18,047 and 19,628 genes in eggplant, pepper and tomato. We found 12,655 genes that were common orthologs (1:1:1) across the three species (Table S5 provide the list of all 1:1 orthologs for all pairs of species and all 1:1:1 orthologs).

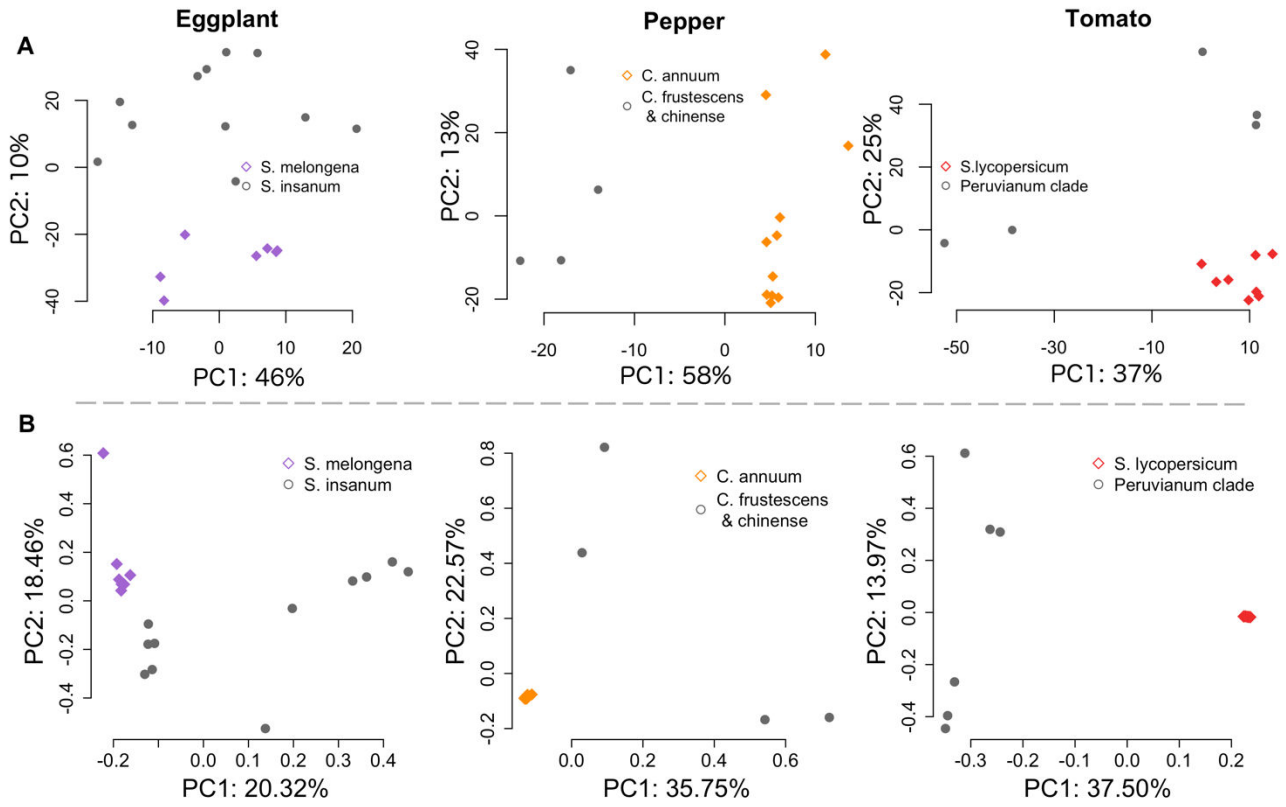



**Figure 1.** Graphical representation of the PCA plots based on: *A.* genetic covariance and *B.* expression level relationships, among all individuals of crop (colored diamonds) and wild (grey circle) individuals for each species. Each dot represents an accession.

## Identification of genetic diversity shifts

From the SNPs, we assessed the genetic distance separating the individuals in each species by performing a PCA (figure 1*A*). The domesticated populations of the three species clustered together which means they have high similarities in their genotypes and less diversity than the accessions from the wild populations that presented greater dispersion within their populations.

The genome-wide nucleotide diversity difference between crop and wild populations was significant for the three species (p-value $9.66\times10^{-06}$ with the Welsh test in eggplant and p-values $<2.2\times10^{-16}$ in both pepper and tomato), with $\pi$ estimates of $4.14\times10^{-04}$ and $6.52\times10^{-04}$ for crop and wild eggplant, respectively, $6.40\times10^{-04}$ and $2.61\times10^{-03}$ for crop and wild pepper and $2.20\times10^{-04}$ and $2.76\times10^{-03}$ for crop and wild tomato. The mean of the genome-wide nucleotide diversity estimates and the results of the statistical tests are detailed in Table S6. Overall, we observed a reduction of nucleotide diversity in crop populations compared to their wild counterparts at the genome-wide scale, in the three species that were significant according to test of Kolmogorov-Smirnov ($<2.2\times10^{-16}$).

The eggplant experienced a decrease in nucleotide diversity that was similar for all chromosomes, but in pepper and tomato, some chromosomes showed significant differences. In pepper, chromosomes 9, 10 and 11 had significantly different regressions of diversity shift according to their expression than chromosome 1 (p-values between 0.0135 and 0.0196). In tomato, chromosomes 6 and 9 were significantly different from chromosome 1 (p-value: 0.0289 and $4.21\times10^{-06}$). The figure S1 is providing a genome-wide representation of the smoothed nucleotide diversity of each species showing its variation.

We plotted the nucleotide diversity in the crop *vs* wild $\pi$ estimates in each species (figure 2) and focused on the shifted genes. In eggplant, we detected 185 shifted genes in the crop-diverse group (i.e. only polymorphic in the crop population) and 369 genes in the wild-diverse group (i.e. only polymorphic in the wild population), 247 and 520 genes in the crop- and wild-diverse groups in pepper, and 64 and 605 genes in the crop- and wild-diverse groups in the tomato, respectively.

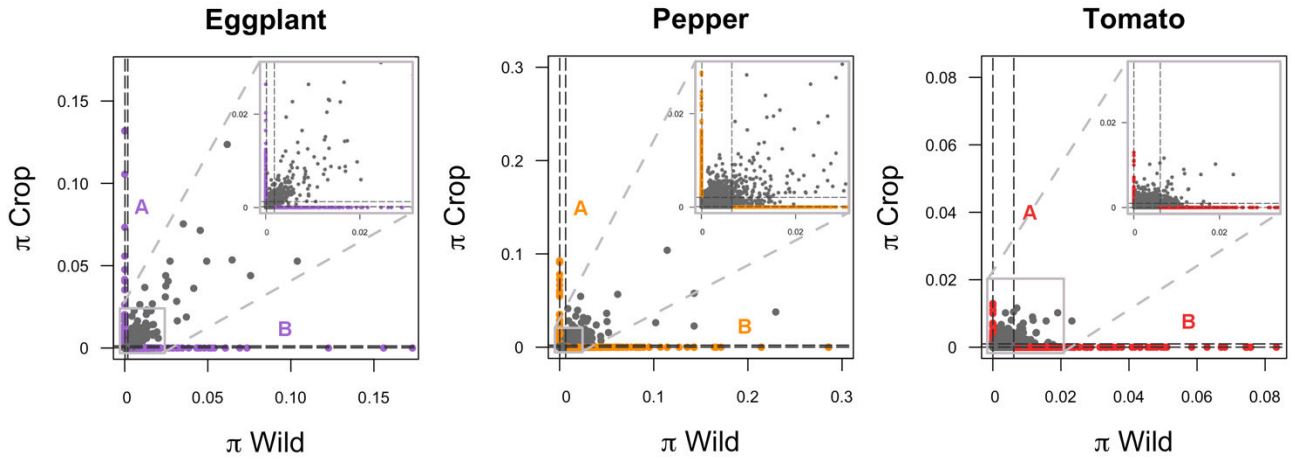



**Figure 2.** Distribution of the nucleotide diversity between the crop population and their wild relatives for each gene. The dots represent the $\pi_{CROP}$ plotted against the $\pi_{WILD}$ for each gene. The colored dots represent the genes that are part of the crop-diverse group (A: top-left) and the wild-diverse group (B: bottom-right).

## Identification of differentially expressed genes

A total of 33,209 CDS showed expression levels (96.5% of the total known CDS) in eggplant, 34,610 CDS (97.9%) in pepper and 34,297 CDS (95.9%) in tomato. In each species, the filtering of paralogous genes reduced the data sets to 17,545 CDS in eggplant, 18,047 CDS in pepper and 19,628 CDS in tomato that we used for the subsequent analyses. Overall within these three sets of expressed genes, the mean of raw read counts per gene ranged from 5.17 to 134,700 (mean = 1,364) in eggplant, from 5.6 to 160,000 (mean = 1,360) in pepper and from 0.85 to 131,400 (mean = 1,041) in tomato.

The PCA analysis, performed on the transformed normalized gene expression levels of each accession (figure 1*B*), showed a clear separation between the crop and the wild populations for each species. After FDR adjustment, the DEG analysis revealed 8,344 DEGs between populations in eggplant (47.6% of the total filtered expressed genes) with an $LFC_{[CROP\ :\ WILD]}$ ranging from -7.02 to 6.04, 987 DEGs between populations in pepper (4.5%) with an LFC ranging from -5.49 to 4.25 and 4,948 DEGs between populations in tomato (25.2%) with an LFC ranging from -5.78 to 6.89 (Table S7). Additionally, for each species, we detected the up- and down-regulated genes (ratio of expression $CROP/WILD$) with 3,924 up- and 4,420 down-regulated DEGs in eggplant, 561 up- and 426 down-regulated DEGs in pepper and 2,170 up- and 2,778 down-regulated DEGs in tomato. Within

the set of 12,655 orthologous genes across the three species, 43 and 48 were systematically up- and down-regulated DEGs.

### Annotations and enrichment analyses

| GO category | Over represented p-value | Number in Group | Number in gene space | Term |
|---|---|---|---|---|
| Crop-diverse | | | | |
| 16480 | $3{,}21\times10^{-03}$ | 1 | 1 | negative regulation of transcription from RNA polymerase III promoter** |
| 6413 | $3{,}40\times10^{-02}$ | 1 | 11 | translational initiation |
| 5992 | $4{,}84\times10^{-02}$ | 1 | 14 | trehalose biosynthetic process |
| Wild-diverse | | | | |
| 3333 | $2{,}48\times10^{-02}$ | 1 | 16 | amino acid transmembrane transport* |
| 9690 | $3{,}46\times10^{-04}$ | 2 | 3 | cytokinin metabolic process** |
| 6817 | $1{,}08\times10^{-02}$ | 1 | 1 | phosphate ion transport |
| 30001 | $1{,}34\times10^{-02}$ | 3 | 46 | metal ion transport |
| 9435 | $2{,}15\times10^{-02}$ | 1 | 2 | NAD biosynthetic process |
| 6471 | $2{,}16\times10^{-02}$ | 1 | 2 | protein ADP-ribosylation |
| 9733 | $3{,}19\times10^{-02}$ | 2 | 26 | response to auxin |
| 55114 | $3{,}85\times10^{-02}$ | 11 | 555 | oxidation-reduction process |
| DEG down | | | | |
| 6597 | $1{,}12\times10^{-02}$ | 1 | 3 | spermine biosynthetic process* |
| 8295 | $1{,}12\times10^{-02}$ | 1 | 3 | spermidine biosynthetic process |
| 30001 | $1{,}38\times10^{-02}$ | 2 | 46 | metal ion transport |
| 55085 | $1{,}91\times10^{-02}$ | 4 | 267 | transmembrane transport |
| DEG up | | | | |
| 6886 | $3{,}25\times10^{-03}$ | 3 | 42 | intracellular protein transport* |
| 6817 | $6{,}68\times10^{-03}$ | 1 | 1 | phosphate ion transport |
| 6098 | $1{,}32\times10^{-02}$ | 1 | 2 | pentose-phosphate shunt |
| 9733 | $1{,}45\times10^{-02}$ | 2 | 26 | response to auxin |
| 6571 | $2{,}12\times10^{-02}$ | 1 | 3 | tyrosine biosynthetic process |
| 48193 | $2{,}71\times10^{-02}$ | 1 | 4 | Golgi vesicle transport |
| 42545 | $3{,}00\times10^{-02}$ | 2 | 38 | cell wall modification |
| 6694 | $4{,}63\times10^{-02}$ | 1 | 7 | steroid biosynthetic process |
| 6270 | $4{,}90\times10^{-02}$ | 1 | 7 | DNA replication initiation |

*Common to the 3 species 1:1:1

**Common to at least 2 species 1:1

**Table 1:** Gene ontology enrichment analyses results for orthologs of 2 or 3 of our species. The group crop-diverse represents the genes more diverse in the crop population and the wild-diverse the genes more diverse in the wild population. The DEGs UP represent the up-regulated genes in crop populations and the DEGs DOWN the one down-regulated.

The Interproscan gene ontology annotation procedure retrieved a total of 81,698 GOs for eggplant, 89,072 GOs for pepper and 85,606 GOs for tomato. When selecting only the Pfam library,

we obtained GO annotations for 18,283 genes (55.05% of the total reference CDS) for eggplant, 17,695 (51.12%) for pepper and 18,093 (50.05%) for tomato. From these GOs, we tested for any significant enrichment in biological processes within (i) the shifted genes (genes that experienced a major loss or gain in nucleotide diversity) and (ii) the DEGs (up- and down-regulated being tested separately).

Firstly, for the shifted genes, at $p<0.01$ threshold, we found only two over-represented GOs in the wild-diverse group of pepper (response to auxin and microtubule-based process) and two GOs in the crop-diverse group of tomato (GO: response to wounding and photosynthesis, light reaction) (For all results at 0.05 significance threshold see figure 3) (extended results are detailed in Table S8).

Secondly, we separately tested the enrichment of the down- and up-regulated sets of DEG for each of the three species (Table S9 for the complete list of GOs per condition). At $p<0.01$ threshold, nine GOs were associated with the down-regulated genes (e.g. DNA replication and cellular regulations, microtubule-based movement and protein modifications) and associated with 13 GOs in up-regulated genes in *S. melongena* (e.g. protein modifications, response to auxin and oxidation-reduction process). In pepper, five GOs were associated with the down-regulated DEG (e.g. cellular transport and oxidation-reduction process) and 10 and eight GOs respectively in down- and up-regulated DEG in *S. lycopersicum* (e.g. down- translation and protein changes; up- oxidation-reduction process and response to auxin).

Then, for the sets of 1:1 (across two species) and 1:1:1 (across three species) orthologs, the enrichment tests revealed that the 1:1 orthologs in the crop-diverse group (n=41) were enriched for three GOs and the 1:1 orthologs (N=147) in the wild-diverse group were enriched in seven GOs. However, no significant enrichment associated with the only one 1:1:1 ortholog gene of the crop-diverse group, while one over-represented GO was assigned to the 1:1:1 ortholog genes (N=17) in the wild-diverse group. Finally, when testing for any BP enrichment within the 43 orthologous up-regulated DEG across the three species, we found four over represented GOs. For the 48 orthologs down-regulated DEG, nine GOs were significantly over-represented. Within the set of 1:1 ortholog genes (across two species), enrichment analyses revealed 10 GOs over-represented in the up-regulated DEGs (mostly linked to translational elongation and terpenoid biosynthetic process) and 13 GOs over-represented in the down-regulated DEGs. Figure 3 shows the numbers of shared orthologs 1:1 and 1:1:1 for the subcategories of shifted genes from the group crop- and wild-diverse

and for the DEGs up- and down-regulated and all GO categories are listed in Table 1. All GO terms are represented in the figure S2.

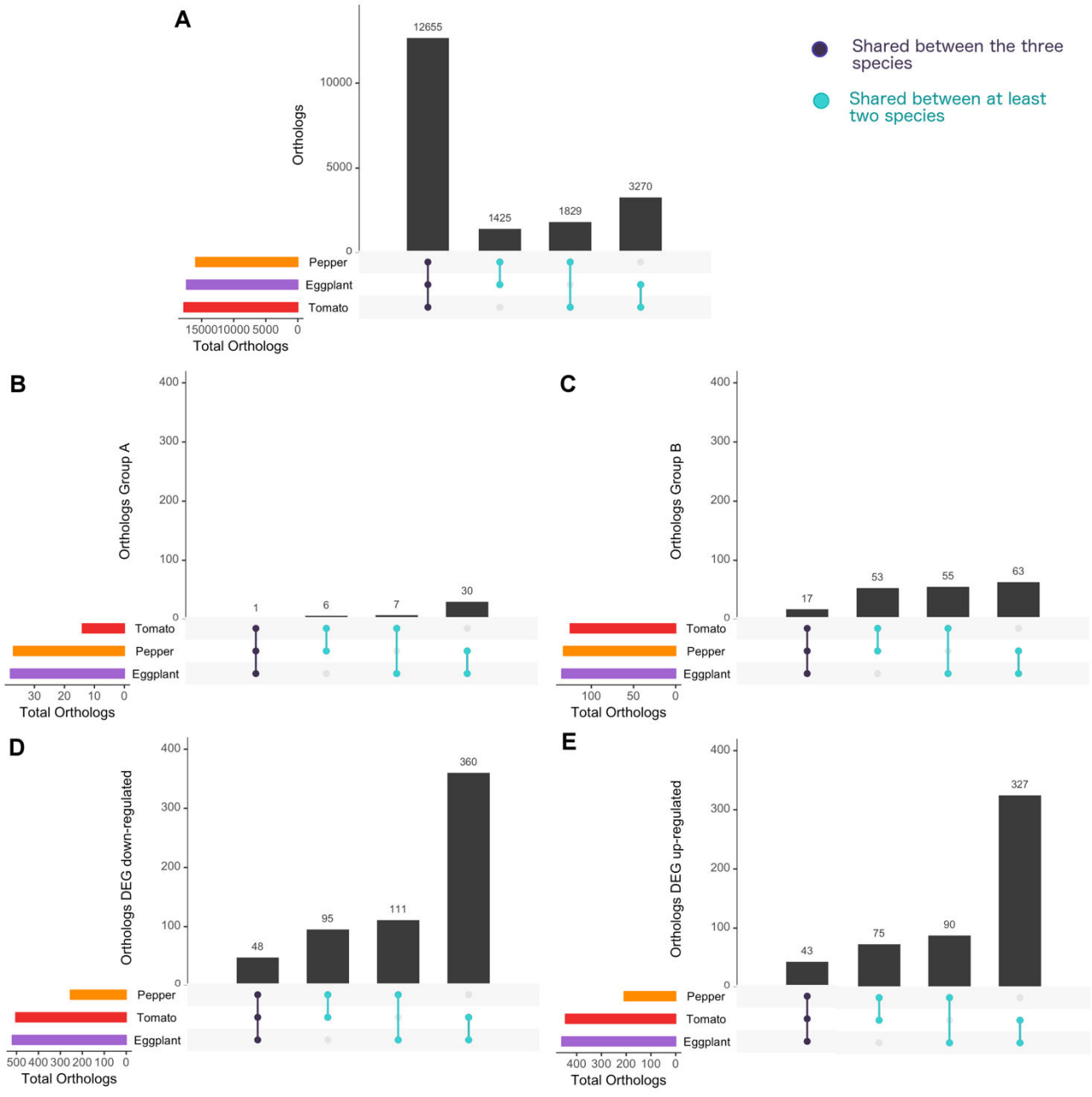



**Figure 3.** Numeric results of the orthology analyses between the three species of *Solanaceae*. *A* is a global orthology analysis on the filtered genes of the three species. *B* and *C* are both the results of the orthology analysis on the nucleotide diversity shifted genes group A (crop-diverse) and group B (wild-diverse). *D* and *E* represent the results of orthology analysis on DEG down- and up-regulated in crop populations.

## Statistical analyses

We used a pairwise Fisher's exact to reveal no significant difference in the number of DEG carried per chromosome in pepper and tomato. In eggplant though, the chromosomes three and five showed a significant higher proportion of DEG compared to the proportion of DEG in the genome (p-value=$5.43\times10^{-02}$ and $1.92\times10^{-02}$), the chromosome three having more down-regulated DEGs and the chromosome five more up-regulated DEGs, but the chromosome seven presented a lower proportion of up-regulated DEGs (p-value=$2.69\times10^{-03}$).

Similarly, the proportion of shifted genes per chromosome compared to the total genome proportion in crop-diverse group (proxy for recent selective pressure in the wild population but diversified in the crop population) was not significant for all the chromosomes in pepper and tomato. In eggplant, the chromosome eight showed a discrepancy of shifted genes (p-value=$7.11\times10^{-03}$). However, the proportion of shifted genes from the wild-diverse group (proxy for directional positive selection in the crop population and relaxation in the wild population) was significantly different for the chromosome six (p-value=$3.66\times10^{-02}$) in eggplant with a higher proportion of selected genes compared to the proportion of shifted genes in the genome, while the chromosome two (p-value=$3.72\times10^{-03}$) in pepper and chromosomes four and nine (p-values=$9.01\times10^{-03}$ and $4.87\times10^{-02}$) in tomato showed a significant lower proportion (the results from the Fisher's exact tests are detailed in the Table S10). At $p<0.1$ threshold, we noted that the proportion of shifted genes from the wild-diverse group was significantly higher in eggplant and pepper only for chromosome nine, while significantly lower in tomato.

The figure 4 provides the graphical representations of the $\Delta\pi$ ($\pi_{WILD}$ - $\pi_{CROP}$) and the LFC for each gene, for the three species. To better understand the relationship between the gene expression levels and the nucleotide diversity, we performed generalized linear models (GLM) for each species considering the chromosome effects. Thus, for each pair of populations, the $\Delta\pi$ of each gene was modeled as a dependent variable with the LFC of the differences in gene expression levels (ratio of gene expression levels in crop over gene expression levels in wild) and the chromosomes as predictor variables (see model (i) in the materials and methods).

For the $\Delta\pi$ as dependent variable, we observed significant chromosome predictor variables: in pepper, chromosomes nine, 10 and 11 showed significant differences in regression coefficients (p-value<0.05) compared to chromosome one (used as reference). In tomato, we found two chromosomes with significant differences in regression coefficients (chromosome six, p-value=0.0289 and chromosome nine, p-value=$4.21\times10^{-06}$). In eggplant, the regression coefficients were not significantly different between chromosomes. When these differences in regression coefficients were considered, the model detected that the LFC was a significant predictor variable for each of the three species (p-value=$6.64\times10^{-15}$ for the eggplant and p-value <$2\times10^{-16}$ in pepper and tomato). The detailed p-value outputs are listed in Table S11.

In the reverse GLM model (ii), we aimed to detect if the '$\Delta\pi + Chromosome$' was a predictor variable of the dependent variable LFC, in each species. We detected significant differences in regression coefficients in pepper for the chromosomes four, five, 12 (p-value<0.05) and chromosomes seven and 10 (p-value<0.01); in tomato for the chromosomes six, 10 (p-value<0.05) and 12 (p-value<0.01) but no significant differences between chromosomes in the eggplant. The chromosome effects considered, we found that $\Delta\pi$ was a significant predictor variable in eggplant (p-value=0.034), pepper and tomato (p-value<$2\times10^{-16}$).

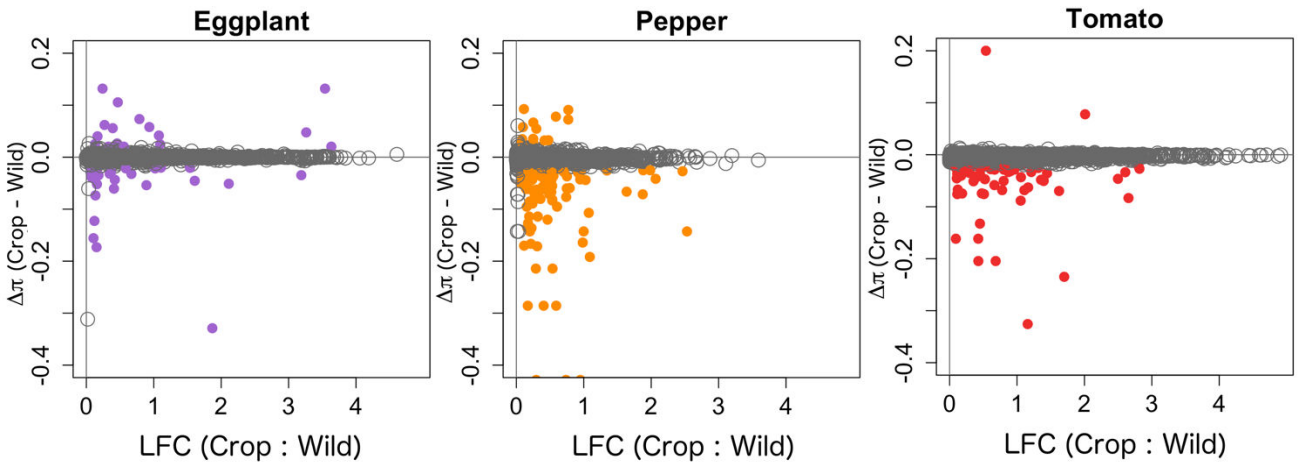



**Figure 4.** Distribution of the genes according to their loss in diversity and to their expression level changes between the crop and the wild population.

## 4. Discussion

Our comparative analysis of the three *Solanaceae* species provides an in-depth view of the changes induced by domestication and modern breeding on gene expression levels and nucleotide diversity patterns between crop and wild populations. We tested for both the trans-specific signatures of selection and species-specific signatures by taking advantages of the high level of synteny between the species genomes of this family. While our transcriptomic analyses confirmed the changes in expression of numerous domestication related genes, the novelty of our study is the highlight of the convergence of domestication footprints acting on both nucleotide diversity and gene expression.

For the three species, we detected genes showing directional selection in the wild (crop-diverse group) or in the crop (wild-diverse group) and examined in details the biological functions associated to them. The proportion of genes from the crop-diverse group was only significantly different among the chromosomes in eggplant, where the chromosome eight showed a lower proportion of genes under diversifying selection compared to the other chromosomes. This lower proportion can be interpreted as the consequence of selective sweeps that occurred during introgression events to confer the monogenic resistances to fusarium and bacterial wilt in this species only (Mutlu et al. 2008; Mutegi et al. 2015). We detected a higher proportion of genes under directional selection in the crop (wild-diverse group) located on the chromosome nine in eggplant and pepper. This stronger selection corroborates with the presence of resistance genes located on this chromosome as previously shown by synteny in tomato (Kortstee et al. 2007; Verlaan et al. 2013; Lin et al. 2014; Zhu et al. 2018). At the opposite, on chromosome nine in tomato, the well described introgression from the wild tobacco mosaic virus resistance to the crop buffers the detection selected in crop (Ohmori et al. 1995).

The gene ontology enrichment analysis allowed the identification, in the three species, of a signature of selective sweep on the circadian clock regulation as previously discovered in tomato with the light-conditional clock deceleration in crop plants (Müller et al. 2018). We found a large number of genes that are associated with general categories such as light and photosynthesis, including genes that experienced diversifying selection across the three crop (crop-diverse group) and genes that experienced stabilizing selection in the pepper crop only (wild-diverse group).

We observed a strong selection acting on genes related to response to wounding, response to biotic stimulus, defense response to bacterium and fungus (i.e. genes of wild-diverse group) in eggplant and tomato but not in pepper. The response to wounding has long been shown in tomato to activate plant defense mechanisms against biotic or abiotic stresses (Conconi et al. 1996; Orozco-Cardenas et al. 2001; Chico et al. 2002). In addition, it has been recently shown that divergence in cis-regulatory sites and, subsequently, transcription factor binding specificity contribute to stress-responsive expression divergence, particularly between wild and domesticated species of tomato (Liu et al. 2018). The defense responses to biotic stresses have been targeted by breeding selection in response to diseases in crop species (Barchi et al. 2011; Verlaan et al. 2013) and clearly identified as driver of transcriptional variation among crop species such as in tomato (Koenig et al. 2013). At the opposite, these enriched GO categories related to defense response to bacteria and fungi are under diversifying selection in the crop pepper population.

Additionally, across the three species, 13 GO categories related to translation were enriched only in the crop population selected genes (wild-diverse group). This enrichment suggests that domestication impacted entire pathways of translation regulation in a similar manner as it negatively regulated the biosynthesis pathway of antinutritional alkaloids in tomato and potato (Itkin et al. 2013).

To test whether we could identify incipient domestication at the mRNA transcript level, we compared patterns of gene expression in the crop and the wild population. Evidences are accumulating that domestication rewires gene expression levels of many crop populations such as cotton (Rapp et al. 2010). maize (Hufford et al. 2013), common bean (Bellucci et al. 2014) and tomato (Koenig et al. 2013; Sauvage et al. 2017). For each of the three *Solanaceae* species, the gene expression levels of all the expressed genes clearly split into two groups, composed of the crop and the wild population. With various percentage of DEG, with 4% in pepper (808 genes), 14% in tomato (2511) and 25% in eggplant (8344), domestication and subsequent selection have profoundly altered the transcriptional landscapes in these three species but at a variable degree. However, these percentages should be cautiously interpreted, as the nucleotide divergence is relatively different between the crop and the wild population within these three species.

In parallel to the enriched categories of genes selected in the crop populations, we found 20 biological processes across the three species whose GO were related to translation mechanism and

for which genes had rewired their levels of gene expression in the crop population. Both eggplant and tomato had the translation GO category enriched with respectively 29.8% and 27.2% of genes from this specific category that were down-regulated. The up-regulated genes in the pepper and tomato crop population were enriched for genes related to tRNA splicing and production, both involved in translation mechanisms recruiting the amino acid and complementing the mRNA to initiate translation (Gruissem 1989). One of the categories that recurrently appeared enriched across the three species (both in the selected genes and the DEG genes) is related to the microtubule-based movement. This category might be affected in crop populations as microtubules are known to be involved in mitoses processes, that could impact any cell developmental processes from the plan vigor to the fruit growth regulation by cell expansion in tomato (Verbelen et al. 2001; Musseau et al. 2017). Knowing phenotypic trait selected during the domestication, these over-represented GO confirm that domestication has imprinted the genome of crop populations both with mutations in coding regions and with modification of gene expression of related genes.

The heterogeneity in the regression coefficients supported the chromosomal differences in nucleotide diversity between crop and wild populations ($\Delta\pi$) across the three species. Once these chromosomal differences fixed in the model, we detected a significative correlation between the expressed genes LFC variation and the nucleotide diversity variation. Using the generalized linear model, we tested whether diversity loss and therefore selection induced by domestication was correlated with expression changes. We found that most of the genes experiencing diversity loss are not experiencing strong changes in expression, though the few that have correlation between this selection and the shifts in expression level change significantly the regression coefficient of the GLM. Therefore, the changes due to domestication are impacting both expression and diversity level and these changes when correlated are the proof a common selection on metabolism variability by controlling the nucleotide changes and the gene expression both at the same time. It corroborates with recent studies on wild and domesticated tomato that show the regulatory selection in wound-responsive genes through cis-regulatory components (Liu et al. 2018). We suggest that nucleotide diversity and gene expression levels diversity evolved in correlation under the selection induced by the domestication and modern breeding. At the molecular level, the underlying mechanisms at play might be related to the alternative splicing that has been shown in Sorghum as a main driver of the canalization of gene expression in this species (Ranwez et al. 2017). However, this hypothesis and its convergence across the *Solanaceae* have to be tested.

From the genetic and the transcriptomic comparative analyses, we evaluated the convergence in the genomic footprints of domestication and breeding, by detecting the 1:1 orthologs across two species and the 1:1:1 orthologs across three species that were unique in each genome (no duplicates) and therefore very conserved. Thus, the results presented here are restricted to genes that have not experienced gene duplication events since the divergence of our species. Any common changes occurring on these conserved genes stresses the role of selection acting on nucleotide diversity and changes in gene expression levels, specifically induced by domestication and modern breeding. Across the crop populations, the surprising yet strongly targeted general category is related to translation as inferred by the DEG and nucleotide diversity analyses.

From the orthologous analyses, we could observe a greater number of genes under positive selection than under a relaxed selection in both the 1:1 and the 1:1:1 orthologs identified (44 in crop-diverse group; 188 in the wild-diverse group). This proportion of genes under selection (188:232, 81%) supports the hypothesis that purifying selection played a major convergent role in shaping the patterns of nucleotide diversity.

We found that these *Solanaceae* crops displayed genes under diversifying selection associated with functional categories related to negative regulation of transcription. The convergence of this accumulation of polymorphism is counter-intuitive but documented as a direct evidence of the cost of domestication (non-removal of slightly deleterious alleles by purifying selection) on transcriptional regulatory elements (Swinnen et al. 2016). Similarly, the biosynthetic process of trehalose evidenced the cost of domestication with enrichment in genes with higher diversity in crop. This sugar is involved in tolerance to abiotic stress, which is less prominent in crop fields than in the wild (Cortina and Culiáñez-Macià 2005).

The orthologs selected in crop populations were enriched in GO categories related to the domestication syndrome phenotypes. The response to auxin drives the fruit development (De Jong et al. 2009). The cytokinin metabolism regulates leaf and plant growth, therefore the resources to control fruit ripening in tomato (Mapelli 1981; Shani et al. 2010; Greco et al. 2012). The phosphate transport modulates the phosphorus as a major macronutrient limiting plant growth (Clarkson and Scattergood 1982; Daram et al. 1998). These results show that fruit and plant growth related traits were preferably selected in our species.

Between the list of DEG established in each species, we searched for orthologous genes that were up or down-regulated to provide an additional proof of selective convergence across the species. Within the orthologous up-regulated genes, one of the over-represented GO categories, the response to auxin, is related to fruit development in tomato (De Jong et al. 2009). Not only the genes of this specific category are differentially expressed, but they show strong signatures of selection as well. Another GO category is related to the cell wall modification that acts upon the tomato cell expansion and fruit ripening (Rose and Bennett 1999). Both of these previous categories support the convergent tuning of biological functions involved in fruit development.

When focusing on the orthologous down-regulated genes, four GO categories were over-represented across the three species: two are general functional categories related to the metal ion transport and the transmembrane transport, limiting the interpretation. But both the spermine and the spermidine biosynthetic processes category are over-represented, acting as growth hormone (Fromm 1997) and playing a key role in the regulation of abiotic stresses in plant (Gill and Tuteja 2010).

Our study uncovered the major molecular consequences of the domestication and modern breeding improvement. Across the three species, the magnitude of differential expression levels revealed a convergent rewiring of genes during the domestication that are in congruence with the ones targeted by selection. In addition, the expressed genes LFC variation was significantly correlated with the nucleotide diversity variation, in the three species. Our study of the Solanaceae family is at the edge of the potential of evolutionary transcriptomics to deepen our knowledge about the molecular consequences of domestication. Similar studies have been proposed to demonstrate convergent signatures in perennial plants (Wu et al. 2018) and in mammals (Alberto et al. 2018).

## Acknowledgements

This work was supported by the EU Marie Curie Career Integration grant (FP7-PEOPLE-2011-CIG grant agreement PCIG10-GA-2011-304164) attributed to CS. SA was supported by a PhD fellowship from the French Région PACA and INRA, in partnership with Gautier Semences. Authors thank the Italian Eggplant Genome Consortium, which includes the DISAFA, Plant Genetics and Breeding (University of Torino), the Biotechnology Department (University of Verona), the CREA-ORL in Montanaso Lombardo (LO) and the ENEA in Rome for providing access to the eggplant genome reference. Thanks

to CRB-lég (https://www6.paca.inra.fr/gafl_eng/Vegetables-GRC) for managing and providing the genetic resources, to Marie-Christine Daunay and Alain Palloix (INRA UR1052) for assistance in choosing the biological materiel used, to Muriel Latreille and Sylvain Santoni from the UMR AGAP (INRA Montpellier, France) for their help with RNAseq library preparation, to Jean-Paul Bouchet (INRA UR1052) for his Bioinformatics assistance and Ivan Scotti (INRA URFM) for his critical reading.

## Author contributions

CS planned and designed the study. CS and RD collected the biological samples. RD conducted the RNAseq libraries preparation. SA performed the analyses and produced the figures under the supervision of CS. SA and CS wrote the manuscript. MC supervised the study and the writing.

## Supplementary Figures

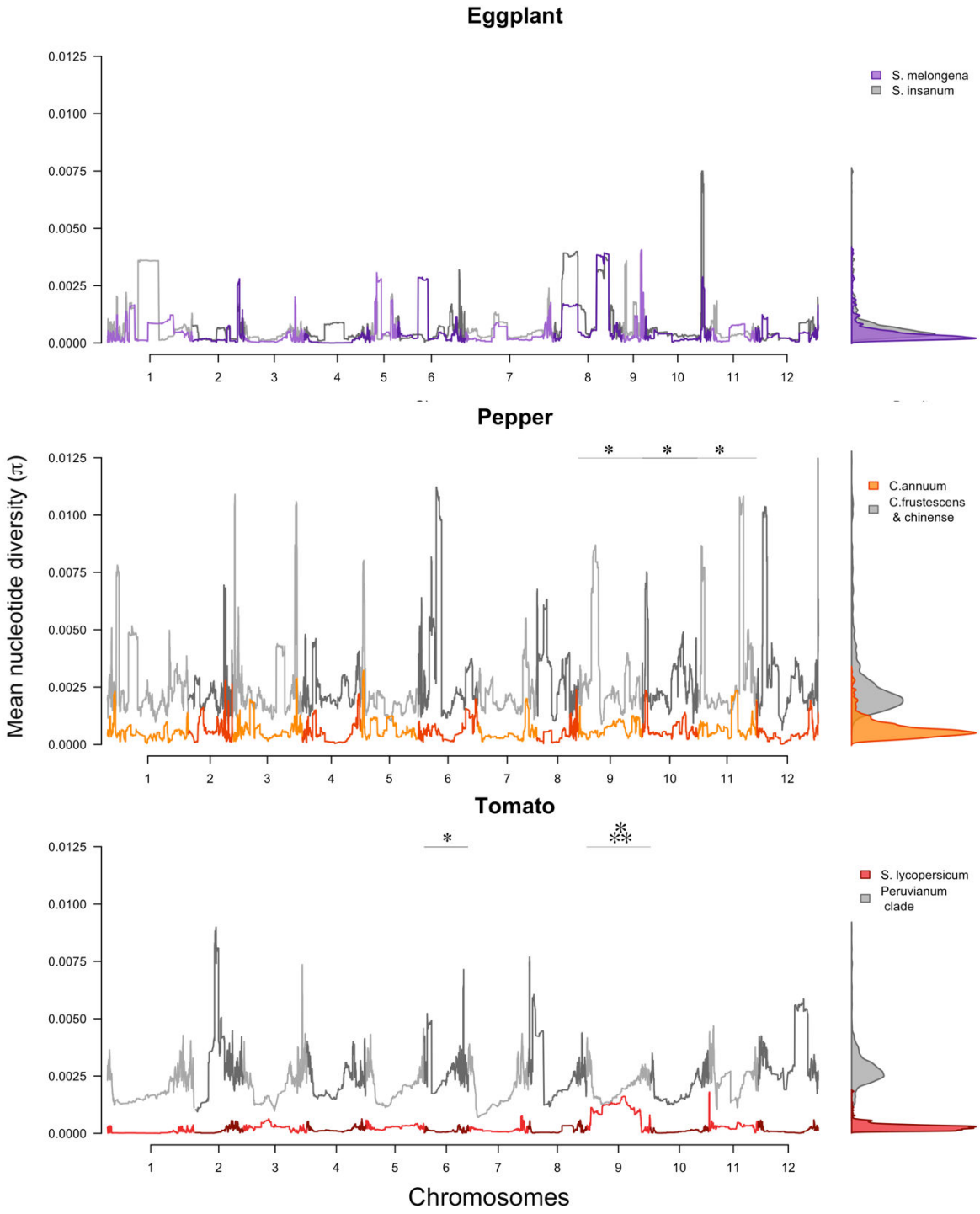



**Figure S1.** Genome wide total nucleotide diversity. ($\pi$ – estimated using sliding windows of 50 genes for each chromosome) the crop (colored) and wild population (grey) in each species. The asterisks represent the p-value significance of the generalized linear model test on the difference between $\pi_{WILD}$ and $\pi_{CROP}$ across the chromosomes. On the right side, the gene mean $\pi$ values are plotted for each population.

**Figure S2.** Gene ontology results according scaled to 1 : p-value, the threshold was set to 0.05 for the enrichment in shifted genes from group A (crop-diverse) and B (wild-diverse) and to 0.01 for the enrichment in DEGs up- and down-regulated.

## Supplementary Tables

The supplementary tables are available in the appendix 3 and are composed of the following tables:

**Table S1:** Detailed data about the studied accessions. The species and the location of origins are listed in separated tables for the three accessions: (a) the eggplant accessions, (b) the pepper accessions and (c) the tomato accessions.

**Table S2:** Raw gene expression for all mapped genes of the studied accessions. The gene expression levels were already filtered for minimum quality. The three species are detailed in separated tables (a) the eggplant accessions, (b) the pepper accessions and (c) the tomato accessions (available on demand or in the published excel file online).

**Table S3:** Mapping summary statistics on mapped, properly paired and singletons of all the studied accessions aligned to their reference genome. The three species are detailed in separated tables (a) the eggplant accessions, (b) the pepper accessions and (c) the tomato accessions.

**Table S4:** Summary of numeric results of expressed genes and SNPs detected with the variant calling for the three species. All details are given before and after filtering for paralogs.

**Table S5:** List of genes that are orthologs between the three species. (available on demand or in the published excel file online)

**Table S6:** Detailed per chromosome and global mean of nucleotide diversity for both populations of our three species.

**Table S7:** Summary results from the DESEQ analyses that detected up- and down-regulated levels of gene expression in crop population compare to the wild population. The summary is detailed for each of the three species.

**Table S8:** Gene ontology analyses results for the nucleotide diversity shifted genes of the three species. The crop-diverse group A represent the genes more diverse in the crop population and the wild-diverse group B the genes more diverse in the wild population.

**Table S9:** Gene ontology analyses results for DEG down- and up- regulated separately for the three species.

**Table S10:** Detailed results from the Fisher test on distribution of the (a) DEG and (b) shifted crop-diverse genes A and wild-diverse genes B across the different chromosomes of the three species.

**Table S11:** Detailed results from the generalized linear models modeling the regression of $\Delta\pi$ and LFC (i) for the three species, and its reciprocal models (ii).

# GENERAL DISCUSSION

This thesis work aimed at conducting a comparative analysis between three Solanaceae crops and their wild relatives, namely the eggplant, pepper and tomato. Such comparative analysis on more than one duo of crop and wild species has not yet been reported in plant science. By studying the convergence of the domestication process we wanted to highlight the crucial potential of wild relatives as a reservoir of adaptive capacity for agricultural systems. With the use of a wide range of bioinformatic methods and tools, the two scientific papers present the PhD work and the answers to the scientific questions that were stated in the introduction. The papers focused on domestication induced changes that impacted the demographic history, and imprinted the molecular scale through gene expression and nucleotide diversity changes. This general discussion will consist in (a) addressing the relevance of the chosen biological material, (b) identifying the advantages and the limitations of the likelihood method to infer demographic models, (c) demonstrating the power of a comparative analysis between crop and wild populations, and its potential outlook of further studies, (d) using the domestication framework to decipher the evolutionary transcriptomic changes, and finally (e) proposing a critical view on the limitation of the domestication studies when neglecting the implementation of environmental conditions. Finally, a general conclusion and outlook is proposed.

### a. Choice of the biological material: does a subset of accessions represents a species?

Sample choices were made with the available knowledge at the project beginning in 2012. Their limited number was also defined according to the sequencing capacity and costs at the time. Since 2012, many advances were made and the cost of genotyping were greatly reduced offering the opportunity to study a greater number of genomes. Indeed, following the crop tomato genomes (The Tomato Genome Consortium 2012), the wild *S. pennellii* reference genome was sequenced (Bolger et al. 2014a), the crop eggplant *S. melongena* draft genome was proposed in 2014 (Hirakawa et al. 2014) and a new version is currently in progress. The reference genome of crop pepper *C. annuum* was sequenced three times (Kim et al. 2014; Qin et al. 2014; Hulse-Kemp et al. 2018) and the wild relative chiltepin *C. annuum* var. glabriusculum was sequenced too (Qin et al. 2014). These efforts in annotating reference genomes are necessary to provide a reference to potential future comparative genome analyses, such as for the 84 accessions of crop, landraces and wild tomato whose whole-genomes were sequenced to decipher the genetic variation available in the he Solanum clade section *lycopersicum* (Aflitos et al. 2014).

In the current context, the genomic variation within the three species is better understood, but at the time, the samples were expected to represent a wide range of genetic diversity within the wild relatives and modern cultivars. The accessions were sampled in different geographical origins. The variation in phenotypes and the previous genetic analyses, based on SSR, guided this choice (*cf*. Introduction). The analyses rely on a small number of accessions and any hypothesis on the domestication events stand on the assumption that the accession panels are representative of the entire species. What appears to be an issue on the analyzed accessions is that some of them were not well annotated and botanically identified, notably for eggplant and pepper. The current sequencing methods could be of use to improve the quality of sampling by using better molecular markers to decipher the accessions annotations, and determine their belonging species. Despite the use of phenotypes, the genetic markers remain more powerful to distinguish species from one another.

The link between the phenotype and genotype was introduced with Gregor Mendel work that developed the principle of heredity, that was the first theory implying that parental phenotypes were transmitted to the descendants (Mendel 1866). Following this principle, the first phylogenetic trees were performed using phenotypic traits with a phenetic method championed by the average distance method UPGMA (Sokal and Michener 1958) and then with molecular and morphological traits using the cladistic method with the neighbor-joining method using operational taxonomic units (Saitou and Nei 1987). It is only with the neutral theory of molecular evolution, that molecular changes were acknowledged to play a key role for most of the variation within and between species (Kimura 1983). The neutral theory of evolution applies only to the molecular markers when phenotypic evolution results from natural selection. Thereafter, when using genotypic data, different phylogenetic methods were used to decipher the plant family taxonomies but neglected the gene flow between species. Whereas phylogenetic methods assume an almost inexistent migration rate between two species, in the framework of domestication studies, using the crop and the wild species pair, the model-based inferences do. While studying domestication, the crop and the wild species remain closely related. During the domestication process, the gene flow between the two compartments has to be considered. Thus, the development and use of model-based inferences have been a great opportunity to improve the studies on the relation between the crop and the wild relative species.

In eggplant, the first taxonomic studies on wild accessions were performed according to morphological traits but were insufficient to classify all the species (Lester, 1986). The total taxonomy of the clade was deciphered in a recent publication on 42 of the 56 recognized species of the clade

(Aubriot et al. 2016). The crop species experienced several taxonomic changes and was structured in 3 morphoforms (group E, G, H) in a study on crop and wild relatives using cpDNA for phenetic and cladistic methods (Sakata and Lester 1997). The groups were considered as artificial and the wild relative progenitor was suggested in a study on *S. melongena*, *S. incanum* and *S. insanum* from Karihaloo in 1995 (Karihaloo and Gottlieb 1995). The wild progenitor *S. insanum* was recently ascertained, after being long debated, in a review of taxonomy from 2016 (Ranil et al. 2016). This work confirms the close relationship between the crop eggplant *S. melongena* and its wild relative *S. insanum*. The PCA on eggplant accessions represent the gradient of differentiation between the crop and the furthest wild accessions. Such gradient is great opportunity to study the process of domestication but it increases the structure and induces difficulties to perform demographic inferences. Following this observation, the closest wild accessions had to be excluded for the demographic inference analyses.

In tomato, the wild accessions have been phenotypically characterized as soon as they were introduced in Europe, by Tournefort (de Tournefort 1694). The tomato being a model plant was studied thoroughly and its taxonomy was proposed by Müller (Müller, 1940) and Luckwill (Luckwill 1943), Child (Child 1990). On molecular data, phenetic (Miller & Tanksley, 1990) and cladistic (Palmer and Zamir 1982; Spooner et al. 1993) analyses helped deciphering the relationships among wild tomato species. The crop wild progenitor of the tomato is *S. pimpinellifolium* and the relationships in the tomato clade were ascertained by a recent entire RNAseq data set (Peralta and Spooner 2000; Pease et al. 2016). This work confirms the previous analyses and pursues the previous effort to understand the modifications due to domestication in tomato.

In pepper, the choice of the biological material is more discussable as shown by the PCA results that revealed *C. annuum var. glabriusculum* not being a direct wild progenitor species but form a structured species strongly differentiated from the crop species. This observation was supported in the present study by the demographic model that estimated the most likely the split between both species as far as 200,000 years old, which is more consistent with a speciation event than any domestication event. Nevertheless, the demographic inferences were powerful enough to detect the bottleneck that stroke the crop population, and, represents most likely the domestication event. It remains unclear which species is the wild progenitor of the *C. annuum* and an accurate identification of such species would increase the power of demographic inferences to decipher parameters of the pepper domestication events. As for other species, eggplant is hypothesized to have been domesticated in several locations (few domestication centers in India and China). A wider

range of accessions (compared to the present study) would permit the detection of genetic structure within the crop species and thereafter, conduct a more detailed/refined demographic inference on multiple species (Jouganous et al. 2017).

While focusing on the inference analyses, the samples even reduced in number were sufficiently informative to allow the inference of demographic parameters with good confident intervals. Therefore, Increasing the number of accessions sampled would not necessary improve the results on the demographic inferences, a good example was the inference study in Maize that used 60 haplotypes out of the 80 samples because of a too high inbreeding (Li et al. 2017). In the chapter 2, the species were represented by 4 to 10 accessions with ~16k filtered SNPs in eggplant, ~41k filtered SNPs in pepper and ~33k filtered SNPs in tomato. Despite the low number of accessions per populations, the work presented remains consistent with recent analyses. Indeed, ðaði was used in few studies to infer domestication events using ~2,000 filtered SNPs for 16 to 20 accessions per species in rice (Molina et al. 2011), ~31,000 filtered SNPs for 11 to 40 accessions per species in *Brassica rapa* (Li 2017) and ~32,000 filtered SNPs for 60 haplotypes per population in maize (Li 2017).

The number of accessions would improve the analyses up to 20 or 30 accessions per populations but most importantly the analyses could benefit from a better representation of the different subset of modern cultivars, landraces and wild relative species that represent the different stages of domestication.

In this context, the work performed here highlights the importance of genotyping crop wild relatives to detect if they are the real wild progenitor as expected (Zeder 2006). The sampling is important especially in species such as wild pepper that remain difficult to differentiate and for which species annotations, relying on phenotypes only, are not powerful enough to decipher the species structures. Therefore, the sampling is one limiting factor to include while analyzing the results of the comparative analyses. Despite the possible improvement of the analyses, the work presented in this thesis already improve considerably the understanding of domestication in the Solanaceae family. The study of the human impact on domesticated plants remains essential to better understand how human societies managed the crop over time, from the cultivation until the modern breeding stages. To decipher the molecular changes, it is necessary to complement phenotypic data by genotypic data. For methods that use phenotype-genotype interactions such as QTL or GWAS, they are the reflection of major effect genes/QTLs and neglect the polygenetic effects (Korte and Farlow 2013). Therefore, using genomic or transcriptomic data seems to be a real alternative to decipher evolutionary changes of pathways at the molecular level (Langridge and Fleury 2011; Koenig et al. 2013).

### b. Advantages and limitations of the likelihood *vs* Bayesian methods for inferring demographic models.

The demographic inferences performed on the polymorphisms differentiating our crop species from their wild relative progenitors were based on the maximum likelihood approach, namely ðaði (Gutenkunst et al. 2009). Such method had been proven to be efficient in case of deciphering the domestication process in Asian rice (Molina et al. 2011) and in cucumber (Qi et al. 2013), where demographic parameters were estimated with ðaði, notably, the population effective size, the migration between both species, and the duration of the bottleneck and following demographic events. One could argue that an approximate Bayesian computation would have been more suitable for the more complex models (Cubry and Vigouroux 2018), but ðaði allows the use of two populations and requires less computational time. Moreover, both approaches are model-based inference and they are the best methods to understand the origin and spread of our domesticated species. Using the comparison of different models was necessary to assess the confidence of each given hypothesis and ascertain the most probable scenario for each species (Gerbault et al. 2014).

Regardless of the approach, the demographic inferences require that the data fit assumptions, e.g. the SNPs need to be independent from each other. In the case of self-crossing plants, some part of the linkage disequilibrium (LD) is not necessarily due to selection but can be caused by selfing (Ellegren and Galtier 2016). In this study the linked loci were pruned to remove the selfing bias and to ensure the SNP details were not redundant. Moreover, by inferring models with heterogeneous migration and heterogeneous effective population size along the chromosomes, we tested for linked selection and differential introgressions (Roux et al. 2013; Sousa and Hey 2013). As both heterogeneities are difficult to dissociate, both categories, i.e. heterogeneous migration and heterogeneous selection, were implemented in our models too, but the data set didn't fit the models with heterogeneity across loci. These results of non-heterogeneity imply that introgressions from wild to crop were not local (on short segments of chromosomes) but globally diffused across the genome in our three species.

In the case of tomato, the crop was known to have several introgressions from *S. peruvianum* located on the chromosome 9. Thus, it was removed for the analyses to ensure it didn't bias the effective population size in the crop. Indeed, most of the introgressions in crop tomato were imported from distant wild species (e.g. *S. habrochaites, S. peruvianum*), but the model infers only introgressions from the wild relative tested (i.e. *S. pimpinellifolium*). Therefore, for the three species,

only a linked selection or differential introgression from the wild tested to the crop, or opposite direction, was tested in our models, and both had poorer likelihood results than homogeneous tests.

These results highlight the global genomic changes due to domestication, it corroborates the previously demonstrated global reduction in nucleotide diversity or the selection on entire genetic pathways and not only on major genes involved in the domestication syndrome.

Additionally, the demographic inferences on domestication are limited to short time scale population differentiation process. The differences in selection are quite strong too, as one of the populations is under natural selection when the other experiences a strong artificial selection. Therefore, following a previous study on Madeiran *Arabidopsis thaliana*, we confirmed our best scenario by assessing the two most probable scenarios, assuming that even with different hypotheses, for two different models, they would have convergent demographic parameters (Fulgione et al. 2018). This comparative method improves the power of the demographic inference by confronting several hypotheses that are the two models that best fit the data. For the interpretation of the results, the models corrupted or with posterior parameters stumbling over boundaries due to overfitting were removed. The corrupted model usually performs better regarding the likelihood but to do so, it infers only part of the data and is consequently unreliable. The best model infers posterior parameters that can be biologically translated according to the mutation rate. The use of a range of mutation rate is necessary here because (i) the mutation rate varies across the genome, and (ii) the mutation rate of the three species, at the genome level, has not been yet estimated. The use of a range of biological estimations prevents an over-interpretation especially in term of dating estimations (Roselius et al. 2005; Lynch 2010).

Moreover, given the hypothesis of multiple domestication events in eggplant, it would be interesting to test a double-founder model as used for the analyses on domesticated rice. Indeed they had three distinct groups of accessions, possibly two domesticated sub-species *Oryza sativa* ssp. *indica* and *O. sativa* ssp. tropical *japonica*, and a wild progenitor *Oryza rufipogon*, and with the use of Bayesian model-based method, they ascertained the two events of domestication (Molina et al. 2011). In the case of eggplant, the sampling would need to be established in separate sub-species to confirm such multi-founder hypothesis. The use of these several sub-species could help deciphering the domestication of crop eggplants, especially for the later stages. Such analyses would be possible with the new methods implementing the software ðaði (Gutenkunst et al. 2009) for multiple species, namely the software named moments (https://bitbucket.org/simongravel/moments) and *momi* (Kamm et al. 2017). Another alternative was the use of a model with secondary contact but the power

is restricted on such short time scale as domestication, and was not relevant for this work. Thus, the sampling and the theoretical methods are a limiting factor for the detection of multiple domestication centers in eggplant with the data set available.

With genomic and not RNAseq data, we could also use pairwise sequentially Markovian coalescent (PSMC)(Li and Durbin 2011) or multiple sequentially Markovian coalescent model (MSMC)(Schiffels and Durbin 2014). These methods translate the estimation of coalescence rate with recombination, into effective population size. The precision of these methods remains low for short-time scales, and new approaches were proposed for short-term inferences, such as the stairway plot (Liu and Fu 2015) or SMC++ (Terhorst et al. 2017). With the complement of both short- and long-term methods, a recent study disentangled African rice history, proving that the 15,000 years old bottleneck of *Oryza glaberrima* (Meyer et al. 2016) and the following long period of low effective population size was present in the crop and in the wild, and correlated with the drying of the Sahara (Cubry et al. 2018). Therefore, the comparative analyses between crop and wild helped understanding that domestication occurred around 2,800 years BCE and that the previously detected bottleneck was indeed a remaining of the crop wild progenitor demography (Cubry et al. 2018).

In this context, the obtaining of genomic data would have improved greatly the precision in estimating the population effective size changes. These approaches combined with the presented demographic inferences would add precious knowledge to decipher the different stages of domestication in our three species with a higher resolution.

### c. Crop and wild comparative analyses are powerful to decipher domestication footprints, the case of the PhD work and further outlook.

The analyses on African rice confirm that comparative analyses are required for deciphering domestication, as the comparison between crop and wild allows to focus on changes due to domestication and not due to natural selection imprinted in the genome since ancestral ages (Cubry et al. 2018). Thus, the analyses performed in this thesis work focus only on the selection footprints present in the crop due to domestication as we assume that evolution did not change much the wild population during the short evolutive time that lasted domestication. To detect selection we focused on the reduction in nucleotide diversity in the crop compared to the wild (i.e. selective sweep)(Smith and Haigh 1974). One concern pointed out by reviewers on the transcriptomic analyses, is that synonymous and non-synonymous SNPs were not differentiated. Knowing that a selective sweep is the change of frequency of neutral alleles at loci that are linked to selected locus, it seems necessary

to differentiate the nucleotide diversity that is neutral (synonymous) to the one that is positively selected (non-synonymous)(Kimura 1983). Three signatures of selective sweep can be detected (Alachiotis and Pavlidis 2018): the local reduction in nucleotide diversity (Smith and Haigh 1974), the shift towards low- and high-frequency derived variants on the SFS (Braverman et al. 1995), and the localized pattern of LD level (Kim and Neilsen 2004). Overall, software focused on one of these signatures at the time, but a recent RAiSD (Raised Accuracy in Sweep Detection) test combines the three statistics and seems promising (Alachiotis and Pavlidis 2018). Despite the global reduction in nucleotide diversity in the crop species highlighted in the chapter 3, the two other methods could improve the detection of selective sweep. The RAiSD methods was developed on genome-wide sequenced data but if it is reliable on the RNAseq data, it could confirm the directional selection observed on the genes targeted by domestication in the chapter 3.

Concerning the SNPs status, the use of an annotation software, such as VEP (McLaren et al. 2016), that perform the analyses to distinct between synonymous and non-synonymous SNPs, require good annotation files such as for tomato (McCarthy et al. 2014). This issue was challenging for both pepper and eggplant, and will need further bioinformatic efforts to improve manually the reference annotation. So far, our analyses focused on the general nucleotide diversity shift between wild and crop species, the results we present are significantly different between both species. Thus when the nucleotide diversity drops to almost null within a crop gene, one can assure that it is a signature of selective sweep, thus positive selection due to domestication (Smith and Haigh 1974). Another statistical focus could have been on analysis of single-marker $F_{st}$ (Lewontin and Krakauer 1973; Chen et al. 2010) calculating the differentiation index between crop and wild and the Tajima's D (Tajima 1996) estimating the selection from the population site frequency spectrum, but unfortunately, the presence of structured populations in our wild species (several species in the wild pepper compartment) did not fulfil the requirement of such statistics. While detecting selective sweeps in crop species, a major concern is the presence of severe bottleneck or introgression that obscure the evolutionary and selective history of a locus (Meyer and Purugganan 2013), without mentioning the strong LD increase due to inbreeding within crop lineages (Ellegren and Galtier 2016).

Concerning the demographic inferences, the data set could, as well, be improved by both ancient DNA and archaeobotanical records. In *Solanaceae*, until now only few published records of archaeobotanical analyses on *Capsicum spp*. are available (Yamaguchi 1983; Duncan et al. 2009; Kraft et al. 2014). The absence of ancient DNA is a pitfall in any demographic inference method as it reduces the possible estimation of an ancestral time point. Despite the high rate of SNPs polarization

produced thanks to the outgroups species, an archeological record would offer a genetic time point. The assumption mentioned earlier that the wild population didn't change from T0 would not be necessary as the archeological record would offer a time point of the real evolutionary state in the past. Such records were used in the Hawaiian petrel populations giving a genetic time point (Welch et al. 2012). Domestication correlates with the cultural development of civilized human populations, and in this work, the use of language and written records served to complement the demographic analyses. Indeed, the oral or written descriptions of crop species are useful thanks to the precise timing of the corresponding archeological records. Translational analyses involving anthropology of the domestication and the dispersion areas could as well improve and complete the global picture on the domestication process as was proven in African rice or pepper (Kraft et al. 2014; Cubry et al. 2018). Therefore, using multiple crop species to perform comparative analyses highlights the convergent aspects of domestication and the species specificity.

Thus a better understanding of the convergence of crop domestication is essential, especially when it comes to produce new domesticates (Stetter et al. 2017). Indeed, neo-domestication are essential to answer societal issues such as energy production or food production, especially in a context of global warming. Producing bioenergy crop is important to avoid the use of fossil fuels, one of the promising species is the *Miscanthus* that can grow in suboptimal land without conflicting with food production. The *Miscanthus* is currently under neo-domestication with a selection that targets yield improvement and other morphological traits (Clifton-brown et al. 2007; Clifton-Brown et al. 2018), genetic diversity (Sang 2011) but as well co-expression patterns (Xing et al. 2018). The neo-domestication of Coffee trees (*Coffea canephora)* was accelerated by gene editing to produce trees that are stress tolerant and resistant to pathogens (Breitler et al. 2018). And in an effort to improve food and nutritional security in Africa (Ofori et al. 2014), two trees producing edible fruits and adapted to the sub-Saharan Africa are currently being domesticated (stage of cultivation). A participatory neo-domestication involving farmers and scientists was performed on the African pear and plum (*Prunus africana* and *Dacryodes edulis*)(Simons and Leakey 2004), by targeting an increase in yield and in fruit traits, such as fruit size (Anegbeh et al. 2005).

### d. Studying evolutionary transcriptomics reveals that domestication induced modifications in different mechanisms of gene expression regulation.

For the RNAseq analyses, all accessions were conserved and grown in controlled conditions reducing the maternal effect (Marshall and Uller 2007). Indeed, responding to the environmental

conditions the plant induces a maternal provisioning that modifies the offspring gene expression (Videvall et al. 2016) via transgenerational epigenetic regulation with embryonic siRNA for example (Autran et al. 2011). Thus, avoiding the maternal effect, by producing plants in similar conditions before to study their offspring, allows to reduce the variation in gene expression level between the accessions.

Moreover, the mix of different tissues in controlled conditions allowed to decipher a large fraction of the expressed genes. This allowed differential expression analyses, but limited our study to the changes at the exome level. It is relevant to study the convergence of the transcriptional regulation and gene structure modification due to domestication, as both mechanisms are involved in the adaptation of crop to domestication. A previous study on tomato had led the path, showing the rewiring of gene co-expression due to domestication (Sauvage et al. 2017). This study greatly inspired the analysis of evolutionary transcriptomics that are presented in this thesis. By studying, both nucleotide diversity shift between crop and wild and differentially expressed genes, the correlation of their modification highlighted the convergence of the regulatory mechanism modification due to domestication.

The ortholog analyses between the genes of the three species revealed that domestication process induced convergent modifications at both gene structure (nucleotide modification) and gene expression levels regardless of the species. But most of all, the results highlighted that biological processes selected during domestication (e.g. domestication syndrome related traits) came with a rewiring of their gene expression. And in addition, the biological processes counter-selected were related to biotic and abiotic stresses tolerance.

In previous studies, while focusing on a set of genes selected during and after domestication in multi-species (e.g. maize, rice, tomato, wheat, pea, etc.), 55 to 63% of these genes were described as transcription factors. Also when half of the mutations are annotated as a loss of function (non-synonymous change), 30 to 43% are annotated as regulatory changes (Doebley et al. 2006; Meyer and Purugganan 2013). These results reinforce the hypothesis that one of the mechanisms of plant evolution relies on the transcriptional regulation. The results presented in the chapter 3 corroborate with this hypothesis as well while highlighting convergent gene expression changes due to domestication. Indeed the modification of *cis*-regulatory elements of transcriptional regulators allows phenotypic changes but reduces the potential pleiotropic impact (Doebley and Lukens 1998).

In 2013, a review highlighted that several phenotypic changes in domesticated plants were potentially due to genomic structural variations, namely changes in copy number variation and in

presence/absence variation (Olsen and Wendel 2013). Following this hypothesis, they also pointed out the major effect of transposable elements within crop species, constituting between 22% and 85% of the total genomic contents of 11 crops (Morrell et al. 2011). The transposon activity has the potential to provide an increased phenotypic diversity, by direct effect through mutagenesis or by indirect effect on the gene expression, the wider range of traits can afterwards be selected during domestication, as for the well-known example of crop maize (Hollister et al. 2011). This type of markers has been neglected so far in the study of Solanaceae domestication, and it would be important to develop further the detection of genomic structural variations in our crop genomes.

Further analyses on intronic regions related to the differentially expressed genes would allow the detection of trans- and cis- regulations as it was done in maize, where they discovered that most of gene expression changes due to domestication were cis- rather than trans-regulated (Lemmon et al. 2014). In tomato, studies found two cis-regulatory mutations regulating the fruit size (Swinnen et al. 2016) but few studies focused on the (cis- and trans-) regulation of domestication phenotypes apart for the regulation of few wound-responsive gene expression (Liu et al. 2018). Therefore, knowing that cis-regulatory mutations impact traits that were selected during domestication, it seems necessary to further study the gene regulation modified during the domestication process.

A recent study in tomato highlighted the importance of epigenetic regulation, in this precise case microRNA regulation, to modulate the expression of targeted genes involved in the biotic stress sensitivity via the production of anthocyanins and $\alpha$-tomatine. Such results underline the importance to deepen the study of epigenetic regulations, especially while focusing on rapid immune responses to cope with threat of pathogens (Chen et al. 2018). Another trait that could be deeper studied is the flowering regulation. Indeed, in rice flowering was proved to be controlled by several chromatin modifiers that are as well mark of epigenetic regulations. This epigenetic regulation allows a complex gene network to integrate environmental signals and plant hormone cues (Albani and Coupland 2010; Shrestha et al. 2014). The epigenetic might be a first mechanism to respond quickly to environmental changes, and deciphering further the epigenetic regulations seems to be promising for adaptive evolution to human-controlled cultivation conditions. With the strong erosion of the genetic diversity, epigenetic diversity emerges as a potential source of phenotypic variation. Plant improvement could rely on it, to increase the crop adaptation to changing environment and to maintain the acquired production performances (Gallusci et al. 2017).

### e. The next improvement to infer domestication might involve the implementing of if the variation in environmental conditions over time.

In these comparisons between crop and wild species, as discussed previously, the consistency of environmental conditions since the beginning of the domestication process is quite a strong assumption. It would be interesting to investigate records and estimate the climatic conditions that possibly changed and induced possible supplementary stresses to the domesticates. This parameter implemented could improve the accuracy in estimating the most likely demographic scenario. During climate change responses, phenotypic plasticity and adaptive evolution contribute to advancing flowering phenology for example with directional selection in Brassicaceae (Anderson et al. 2012). This is consistent with the domestication changes that are related to light sensitivity, such as the postdomestication day-length adaptation that modified the maize and prompted its spread to temperate zones (Hung et al. 2012). This is a good example on how using rapid directional selection could offer possible alternative for future modern breeding. Indeed, recent studies on tomato revealed as well the impact of domestication on the adaptation to day length with the loss of day-length-sensitive flowering (Soyk et al. 2017). This adaptation to day length is required to increase the geographical range of crop cultivation. In this context of world-wide cultivation of the crops, implementing the environmental condition as a fluctuating variable could considerably increase the accuracy of our models.

The range of possible phenotypic acclimation within a crop species is another parameter that could improve the model-based methods. The high phenotypic plasticity can help crop to adapt to new or changing environment at the individual level without imprinting directly the genome, though a selection of specific phenotypes can induce genomic footprints over generations, falling into the protracted theory (Allaby 2010). In maize, a recent paper highlighted the different phenotypes, as commonly accepted, of both crop and wild current accessions when they were grown under the early domestication conditions (Lorant et al. 2017). Therefore, implementing the environmental condition into the modeling could refine our understanding of the domestication process. The ecology reveals itself already essential to decipher domestication centers such as for the *C. annuum*, where they used complementary data spatially located such as the archeology records, the linguistics, the genetic distance analyses, and, including the ecological data corresponding to the paleoclimatic conditions during the mid-Holocene (Kraft et al. 2014). In our analyses we detected in eggplant a bottleneck in both crop and wild populations, this could be the signal of a few years of highly stressed conditions

during a climatic crisis [drought / fires] that would have affected both crop and wild populations imprinting their genomes as was found for the bottleneck in African rice during the drying of the Sahara (Cubry et al. 2018).

Another important aspect of the plant physiology that would need to be implemented to fulfil the domestication model is the metabolomic diversity. Indeed, understanding the metabolomic changes due to domestication could highlight pathways of future improvement. While most of crop breeding focused on genetic diversity so far, new approaches relying on transcriptomic and metabolomic changes offer new opportunities to elicit yield and nutritional traits enhancement (Harrigan et al. 2007). Recently, a metabolomic profile of crop and wild Soybean revealed metabolites that were involved in the tolerance to salt stress in the wild Soybean but had been lost during the domestication (Zhang et al. 2016a). The study of metabolomic profile in tomato was performed recently and highlighted the rewiring of the fruit metabolome (Zhu et al. 2018). These analyses are precursor to metabolome-assisted breeding programs. In this context, such study would have complemented the overview we intended to obtained on the molecular footprints of domestication in the crop of the Solanaceae family.

## General conclusion and outlook

- What was the wild progenitor species of the current crop?

The first question concerning the wild progenitor of each species seemed to be answered when the samples were chosen. With our analyses we can ascertain, once again, that *Solanum melongena* was domesticated from the wild progenitor *S. insanum*, and that *S. pimpinellifolium* is the wild progenitor of *S. lycopersicum*. *Capsicum annuum* var. *glabriusculum*, the supposed wild progenitor of the crop pepper, needs to be further studied as our results ascertain the strong discrepancy of the species. The crop pepper was surely domesticated from a sub-species of *C. annuum* var. *glabriusculum* but the species structure has to be disentangled for further analyses on domestication.

- How much domestication impacted genomes and transcriptomes of crop species?

The study focused on RNAseq, thus on the expressed part of the genome. By comparing crop and wild relative species, in the three Solanaceae species, we could detect considerable changes in nucleotide diversity and in gene expression level due to domestication. The correlation between these changes (genetic diversity and gene expression variation) revealed the convergence of the mechanisms of regulation at the genome and transcriptome scale while adapting to domestication.

Further study on metabolome experiments could lead to a more complete understanding of each level of molecular regulation.

- What were the genes and pathways targeted by selection?

The ortholog study, that revealed common genes targeted by domestication within the three species, revealed a convergence in selection due to domestication. Domestication positively impacted traits that were related to the domestication syndrome while altering pathways involved in stress tolerance and in diseases resistance.

- And finally, what can be retrieved from the wild relative species to improve modern cultivars?

While identifying domestication-target genes and pathways within the crop species, with the comparative analyses, the wild relative species reveals its potential as genetic resource for the recovery of the identified selected traits. The genetic diversity that remains in wild relatives is an opportunity to improve greatly the weaknesses of the crop species or modern cultivars, e.g. to recover disease resistances or environmental stress tolerance.

This work confirms the necessity to conserve wild relatives and landraces in more representative core collections. Especially, in a context where landraces, that were maintained for thousands of years, are slowly disappearing with the rural flight of indigenous populations, such as for indigenous Amerindian populations that were the conservation center of most of the old landraces (Smith et al. 1992).

This thesis focused on the common and divergent features of the crop and wild relative species. The comparative methods on RNAseq were a great opportunity to decipher the changes in expression and in nucleotide polymorphism due to domestication. The results presented here, provide, from the demographic inferences, an estimation of the domestication events duration, and, from the transcriptomic analyses, an overview of the genetic and transcriptomic consequences of the domestication process. Both papers confirm the loss of adaptive diversity and the loss in genetic diversity within crop species.

Overall, these results offer the opportunity to foresee future improvements related to the loss of adaptability genes within crop species that remain in the wild relative species. In the late 20$^{th}$ century (since 1945), more than 30% of the increased crop yields can be attributed to the use of CWRs in plant breeding programs (Pimentel et al. 1997). In this context, the direct implication of this work highlights the necessity to support the conservation of wild relative species in wild locations and in seed stock center. The analyses in domestication changes reveal, notably, traits of interest

remaining in the wild gene pools. The landrace and wild species could be used as part of the reference population for future genome wide analyses to detect regions potentially source of improvement. The detected regions could then be introgressed into modern cultivars to improve their tolerance to stresses and resistances to pathogens. In parallel, epigenetic and metabolomic variations are both sources of phenotypic diversity. Thus, using emerging biotechnology the modification in gene regulation and in metabolic composition could lead to essential yield and nutritional traits improvements for crops (Harrigan et al. 2007; Gallusci et al. 2017). Eventually, the modern breeding efforts would increase considerably phenotypic and genotypic data allowing the use of genomic selection. This method connects the known phenotypes and genotypes, and uses them as prior to model and predict phenotypes from the genotypes (Morrell et al. 2011).

Such work provides a backbone platform to modern breeding programs by providing a list of genes that were communally targeted during domestication in the three Solanaceae species. The convergence of these changes, offers a considerable opportunity to use transversal knowledge to improve crops, for example, using trans-species gene editing (Bastet et al. 2018). Especially when considering the high synteny present within the Solanaceae family and that offers the opportunity to transfer knowledges to other species (Rinaldi et al. 2016)**.**

This thesis work confirms what Darwin suggested more than a hundred years ago already that studying the domestication process has a great potential to better understand artificial selection and convergent evolution as much as to bring valuable insights for improvement and breeding effort.

# BIBLIOGRAPHY


Abberton M, Batley J, Bentley A, et al (2016) Global agricultural intensification during climate change: a role for genomics. Plant Biotechnol J 14:1095–1098. doi: 10.1111/pbi.12467

Abbo S, Pinhasi van-Oss R, Gopher A, et al (2014) Plant domestication versus crop evolution: A conceptual framework for cereals and grain legumes. Trends Plant Sci 19:351–360. doi: 10.1016/j.tplants.2013.12.002

Adetula O (2006) Genetic diversity of Capsicum using random amplified polymorphic DNAs. Afr J Biotechnol 5:120–122

Aflitos S, Schijlen E, de Jong H, et al (2014) Exploring genetic variation in the tomato ( *Solanum* section *Lycopersicon* ) clade by whole-genome sequencing. Plant J 80:136–148. doi: 10.1111/tpj.12616

Aguilar-Melendez A, Morrell PL, Roose ML, Kim S-C (2009) Genetic diversity and structure in semiwild and domesticated chiles (*Capsicum annuum*; Solanaceae) from Mexico. Am J Bot 96:1190–1202. doi: 10.3732/ajb.0800155

Akbar N, Habib A, Ghafoor S, et al (2010) Estimation of genetic diversity in *Capsicum* germplasm Using randomly amplified polymorphic DNA. Asian J Agric Sci 2(2) 53-56, 2010 2:53–56

Alachiotis N, Pavlidis P (2018) RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. Commun Biol 1:79. doi: 10.1038/s42003-018-0085-8

Albani MC, Coupland G (2010) Comparative analysis of flowering in annual and perennial plants. Curr Top Dev Biol 91:323–348. doi: 10.1016/S0070-2153(10)91011-9

Albert E, Segura V, Gricourt J, et al (2016) Association mapping reveals the genetic architecture of tomato response to water deficit: focus on major fruit quality traits. J Exp Bot 67:6413–6430. doi: 10.1093/jxb/erw411

Alberto FJ, Boyer F, Orozco-Terwengel P, et al (2018) Convergent genomic signatures of domestication in sheep and goats. Nat Commun 9:. doi: 10.1038/s41467-018-03206-y

Albrecht E, Escobar M, Chetelat RT (2010) Genetic diversity and population structure in the tomato-like nightshades *Solanum lycopersicoides* and *S. sitiens*. Ann Bot 105:535–554. doi: 10.1093/aob/mcq009

Ali Z, Xu ZL, Zhang DY, et al (2011) Molecular diversity analysis of eggplant (*Solanum melongena*) genetic resources. Genet Mol Res 10:1141–1155. doi: 10.4238/vol10-2gmr1279

Allaby R (2010) Integrating the processes in the evolutionary system of domestication. J Exp Bot 61:935–944. doi: 10.1093/jxb/erp382

Allaby RG, Fuller DQ, Brown TA (2008) The genetic expectations of a protracted model for the origins of domesticated crops. Proc Natl Acad Sci 105:13982–13986. doi: 10.1073/pnas.0803780105

Allen NC, Bagade S, McQueen MB, et al (2008) Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: The SzGene database. Nat Genet 40:827–834. doi: 10.1038/ng.171

Anderson JT, Inouye DW, McKinney AM, et al (2012) Phenotypic plasticity and adaptive evolution contribute to advancing flowering phenology in response to climate change. Proc R Soc B Biol Sci 279:3843–3852. doi: 10.1098/rspb.2012.1051

Andrews J (1993) Diffusion of Mesoamerican food complex to southeastern Europe. Geogr Rev 83:194. doi: 10.2307/215257

Anegbeh PO, Ukafor V, Usoro C, et al (2005) Domestication of *Dacryodes edulis*: 1. Phenotypic variation of fruit traits from 100 trees in southeast Nigeria. New For 29:149–160. doi: 10.1007/s11056-005-0266-4

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Report 9:208–218. doi: 10.1007/BF02672069

Arunyawat U, Stephan W, Städler T (2007) Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. Mol Biol Evol 24:2310–2322. doi: 10.1093/molbev/msm162

Atherton JG, Harris GP (1986) Flowering. In: The Tomato Crop. Springer Netherlands, Dordrecht, pp 167–200

Aubriot X, Knapp S, Syfert MM, et al (2018) Shedding new light on the origin and spread of the brinjal eggplant ( *Solanum melongena* L.) and its wild relatives. Am J Bot 105:1175–1187. doi: 10.1002/ajb2.1133

Aubriot X, Singh P, Knapp S (2016) Tropical Asian species show that the Old World clade of “spiny solanums” (*Solanum* subgenus *Leptostemonum* pro parte: Solanaceae) is not monophyletic. Bot J Linn Soc 181:199–223. doi: 10.1111/boj.12412

Autran D, Baroux C, Raissig MT, et al (2011) Maternal epigenetic pathways control parental contributions to arabidopsis early embryogenesis. Cell 145:707–719. doi: 10.1016/j.cell.2011.04.014

Baek YS, Royer SM, Broz AK, et al (2016) Interspecific reproductive barriers between sympatric populations of wild tomato species ( *Solanum* section *Lycopersicon* ). Am J Bot 103:1964–1978. doi: 10.3732/ajb.1600356

Bai Y, Lindhout P (2007) Domestication and breeding of tomatoes: What have we gained and what can we gain in the future? Ann Bot 100:1085–1094. doi: 10.1093/aob/mcm150

Barchi L, Lanteri S, Portis E, et al (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. BMC Genomics 12:1–9. doi: 10.1186/1471-2164-12-304

Barrero LS, Tanksley SD (2004) Evaluating the genetic basis of multiple-locule fruit in a broad cross section of tomato cultivars. Theor Appl Genet 109:669–679. doi: 10.1007/s00122-004-1676-y

Barton NH, Hewitt GM (1989) Adaptation, speciation and hybrid zones. Nature 341:497

Bastet A, Lederer B, Giovinazzo N, et al (2018) Trans-species synthetic gene design allows resistance pyramiding and broad-spectrum engineering of virus resistance in plants. Plant Biotechnol J 1–13. doi: 10.1111/pbi.12896

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw 67:. doi: 10.18637/jss.v067.i01

Bauchet G, Causse M (2012) Genetic Diversity in Tomato (*Solanum lycopersicum*) and Its Wild Relatives. In: Genetic Diversity in Plants. InTech

Bauchet G, Grenier S, Samson N, et al (2017a) Use of modern tomato breeding germplasm for deciphering the genetic control of agronomical traits by Genome Wide Association study. Theor Appl Genet 130:875–889. doi: 10.1007/s00122-017-2857-9

Bauchet G, Grenier S, Samson N, et al (2017b) Identification of major loci and genomic regions controlling acid and volatile content in tomato fruit: implications for flavor improvement. New Phytol 215:624–641. doi: 10.1111/nph.14615

Beddows I, Reddy A, Kloesges T, Rose LE (2017) Population genomics in wild tomatoes—the interplay of divergence and admixture. Genome Biol Evol 9:3023–3038. doi: 10.1093/gbe/evx224

Bedinger PA, Chetelat RT, McClure B, et al (2011) Interspecific reproductive barriers in the tomato clade: opportunities to decipher mechanisms of reproductive isolation. Sex Plant Reprod 24:171–187. doi: 10.1007/s00497-010-0155-7

Beissinger TM, Wang L, Crosby K, et al (2016) Recent demography drives changes in linked selection across the maize genome. Nat Plants 2:16084. doi: 10.1038/NPLANTS.2016.84

Bellucci E, Bitocchi E, Ferrarini A, et al (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

Benazzo A, Panziera A, Bertorelle G (2015) 4P: Fast computing of population genetics statistics from large DNA polymorphism panels. Ecol Evol 5:172–175. doi: 10.1002/ece3.1261

Bhagirath C, Kadambini G (2009) The development and regulation of Bt brinjal in India (eggplant/aubergine). ISAAA Brief No.38. ISAAA Briefs xii + 102 pp.

Birchler JA, Yao H, Chudalayandi S, et al (2010) Heterosis. Plant Cell Online 22:2105–2112. doi: 10.1105/tpc.110.076133

Blackman BK, Rasmussen DA, Strasburg JL, et al (2011) Contributions of flowering time genes to sunflower domestication and improvement. Genetics 187:271–287. doi: 10.1534/genetics.110.121327

Blanca J, Cañizares J, Cordero L, et al (2012) Variation revealed by SNP genotyping and morphology provides insight into the origin of the tomato. PLoS One 7:e48198. doi: 10.1371/journal.pone.0048198

Blanca J, Montero-Pau J, Sauvage C, et al (2015) Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. BMC Genomics 16:257. doi: 10.1186/s12864-015-1444-1

Bohs L, Olmstead RG (1997) Phylogenetic relationships in *solanum* (solanaceae) based on ndhf sequences. Syst Bot 22:5. doi: 10.2307/2419674

Bolger A, Scossa F, Bolger ME, et al (2014a) The genome of the stress-tolerant wild tomato species *Solanum pennellii.* Nat Genet 46:1034–1038. doi: 10.1038/ng.3046

Bolger AM, Lohse M, Usadel B (2014b) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. doi: 10.1093/bioinformatics/btu170

Bombarely A, Menda N, Tecle IY, et al (2011) The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. Nucleic Acids Res 39:1149–1155. doi: 10.1093/nar/gkq866

Bombarely A, Moser M, Amrad A, et al (2016) Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*. Nat Plants In press:1–9. doi: 10.1038/nplants.2016.74

Böndel KB, Lainer H, Nosenko T, et al (2015) North-south colonization associated with local adaptation of the wild tomato species *Solanum chilense*. Mol Biol Evol 32:2932–2943. doi: 10.1093/molbev/msv166

Borevitz JO, Nordborg M (2003) The impact of genomics on the study of natural variation in *Arabidopsis*. Plant Physiol 132:718–725. doi: 10.1104/pp.103.023549

Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol 12:232. doi: 10.1186/gb-2011-12-10-232

Bradley D, Carpenter R, Copsey L, et al (1996) Control of inflorescence architecture in *Antirrhinum*. Nature 379:791

Bradley D, Ratcliffe O, Vincent C, et al (1997) Inflorescence commitment and architecture in *Arabidopsis*. Science 275:80–83. doi: 10.1126/science.275.5296.80

Brandvain Y, Kenney AM, Flagel L, et al (2014) Speciation and Introgression between *Mimulus nasutus* and *Mimulus guttatus*. PLoS Genet 10:e1004410. doi: 10.1371/journal.pgen.1004410

Braverman JM, Hudson RR, Kaplan NL, et al (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783–796.

Breitler JC, Dechamp E, Campa C, et al (2018) CRISPR/Cas9-mediated efficient targeted mutagenesis has the potential to accelerate the domestication of *Coffea canephora*. Plant Cell Tissue Organ Cult 134:383–394. doi: 10.1007/s11240-018-1429-2

Brown AHD, Marshall DR (1995) A basic sampling strategy: theory and practice. In: Collecting plant genetic diversity: Technical guidelines. pp 75–91

Brozynska M, Furtado A, Henry RJ (2015) Genomics of crop wild relatives: expanding the gene pool for crop improvement. Plant Biotechnol J 14:n/a-n/a. doi: 10.1111/pbi.12454

Brubaker CL, Wendel JF (1994) Reevaluating the Origin of Domesticated Cotton (*Gossypium hirsutum*; *Malvaceae*) Using Nuclear Restriction Fragment Length Polymorphisms (RFLPs). Am J Bot 81:1309. doi: 10.2307/2445407

Caicedo AL, Williamson SH, Hernandez RD, et al (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet 3:1745–1756. doi: 10.1371/journal.pgen.0030163

Canady MA, Meglic V, Chetelat RT (2005) A library of *Solanum lycopersicoides* introgression lines in cultivated tomato. Genome 48:685–697. doi: 10.1139/g05-032

Carputo D, Monti L, Werner JE, Frusciante L (1999) Uses and usefulness of endosperm balance number. TAG Theor Appl Genet 98:478–484. doi: 10.1007/s001220051095

Carrari F, Baxter C, Usadel B, et al (2006) Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. Plant Physiol 142:1380–1396. doi: 10.1104/pp.106.088534

Carrizo García C, Barfuss MHJJ, Sehr EM, et al (2016) Phylogenetic relationships, diversification and expansion of chili peppers (*Capsicum*, Solanaceae). Ann Bot 118:mcw079. doi: 10.1093/aob/mcw079

Causse M, Desplat N, Pascual L, et al (2013) Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. BMC Genomics 14:. doi: 10.1186/1471-2164-14-791

Causse M, Duffe P, Gomez MC, et al (2004) A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. J Exp Bot 55:1671–1685

Causse M, Grandillo S (2016) Gene mapping in tomato. In: Causse M, Giovannoni J, Bouzayen M, Zouine M (eds) The Tomato Genome. Springer, Berlin, Heidelberg, pp 23–37

Chakrabarti M, Zhang N, Sauvage C, et al (2013) A cytochrome P450 regulates a domestication trait in cultivated tomato. Proc Natl Acad Sci 110:17125–17130. doi: 10.1073/pnas.1307313110

Chang CC, Chow CC, Tellier LC, et al (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7. doi: 10.1186/s13742-015-0047-8

Charlesworth D, Wright SI (2001) Breeding systems and genome evolution. Curr Opin Genet Dev 11:685–690. doi: 10.1016/S0959-437X(00)00254-9

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

Chen L, Meng J, He XL, et al (2018) *Solanum lycopersicum* miR1916 targets multiple target genes and negatively regulates the immune response in tomato. Plant Cell Environ pce.13468. doi: 10.1111/pce.13468

Chico JM, Raíces M, Téllez-Iñón MT, Ulloa RM (2002) A calcium-dependent protein kinase is systemically induced upon wounding in tomato plants. Plant Physiol 128:256–270. doi: 10.1104/pp.010649

Child A (1990) A synopsis of *Solanum* subgenus *Potatoe* (G. Don) (d'Arcy) (*Tuberarium* (Dun.) BITTER (s. 1.)). Feddes Repert 101:209–235. doi: 10.1002/fedr.19901010502

Chitwood DH, Kumar R, Headland LR, et al (2013) A quantitative genetic basis for leaf morphology in a set of precisely defined tomato introgression lines. Plant Cell 25:2465–2481. doi: 10.1105/tpc.113.112391

Clark RM, Linton E, Messing J, Doebley JF (2004) Pattern of diversity in the genomic region near the maize domestication gene tb1. Proc Natl Acad Sci 101:700–707. doi: 10.1073/pnas.2237049100

Clarkson DT, Scattergood CB (1982) Growth and phosphate transport in barley and tomato plants during the development of, and recovery from, phosphate-stress. J Exp Bot 33:865–875. doi: 10.1093/jxb/33.5.865

Clement CR (1999) 1492 and the loss of amazonian crop genetic resources. I. The relation between domestication and human population decline. Econ Bot 53:188–202. doi: 10.1007/BF02866498

Clifton-Brown J, Harfouche A, Casler MD, et al (2018) Breeding progress and preparedness for mass-scale deployment of perennial lignocellulosic biomass crops switchgrass, miscanthus, willow and poplar. GCB Bioenergy. doi: 10.1111/gcbb.12566

Clifton-brown JC, Breuer J, Jones MB (2007) Carbon mitigation by the energy crop, *Miscanthus*. Glob Chang Biol 13:2296–2307. doi: 10.1111/j.1365-2486.2007.01438.x

Coffman AJ, Hsieh PH, Gravel S, Gutenkunst RN (2016) Computationally efficient composite likelihood statistics for demographic inference. Mol Biol Evol 33:591–593. doi: 10.1093/molbev/msv255

Conconi A, Miquel M, Browse JA, Ryan CA (1996) Intracellular levels of free linolenic and linoleic acids increase in tomato leaves in response to wounding. Plant Physiol 111:797–803. doi: 10.1104/PP.111.3.797

Cong B, Barrero LS, Tanksley SD (2008) Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. Nat Genet 40:800–804. doi: 10.1038/ng.144

Cornille A, Gladieux P, Smulders MJM, et al (2012) New insight into the history of domesticated apple: Secondary contribution of the European wild apple to the genome of cultivated varieties. PLoS Genet 8:. doi: 10.1371/journal.pgen.1002703

Cortina C, Culiáñez-Macià FA (2005) Tomato abiotic stress enhanced tolerance by trehalose biosynthesis. Plant Sci 169:75–82. doi: 10.1016/j.plantsci.2005.02.026

Coyne JA, Orr HA (2004) Speciation. Sinauer

Crossa J, Pérez P, Hickey J, et al (2013) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (Edinb) 112:48

Croteau R, Kutchan TM, Lewis NG (2000) Secondary metabolites from *Spirotropis longifolia* (DC) Baill and their antifungal activity against human pathogenic fungi. In: Buchanan B, Gruissem W, Jones R (eds) Biochemistry & Molecular Biology of Plants. American Society of Plant Physiologists, pp 1250–1318

Cubry P, Tranchant-Dubreuil C, Thuillet AC, et al (2018) The rise and fall of african rice cultivation revealed by analysis of 246 new genomes. Curr. Biol. 28:2274–2282.e6

Cubry P, Vigouroux Y (2018) Population genomics of crop domestication: Current State and Perspectives. 1–23. doi: 10.1007/13836_2018_48

Dalal A, Rana JS, Kumar A (2017) Ultrasensitive nanosensor for detection of malic acid in tomato as fruit ripening indicator. Food Anal Methods 10:3680–3686. doi: 10.1007/s12161-017-0919-x

Danecek P, Auton A, Abecasis G, et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158. doi: 10.1093/bioinformatics/btr330

Daram P, Brunner S, Persson BL, et al (1998) Functional analysis and cell-specific expression of a phosphate transporter from tomato. Planta 206:225–233. doi: 10.1007/s004250050394

Darwin C (1859) On the origins of species by means of natural selection. Murray, London

Darwin CR (1868) The variation of animals and plants under domestication. In: Murray J (ed) 1st ed. London

Daunay M-C, Laterrot H (2007) Iconography of the Solanaceae from antiquity to the XVII th century: a rich source of information on genetic diversity and uses. John Wiley & Sons, Inc

Davidar P, Snow AA, Rajkumar M, et al (2015) The potential for crop to wild hybridization in eggplant (*Solanum melongena*; Solanaceae) in Southern India. Am J Bot 102:129–139. doi: 10.3732/ajb.1400404

de Candolle A (1886) The origin of cultivated plants. Cambridge University Press, Cambridge

De Jong M, Mariani C, Vriezen WH (2009) The role of auxin and gibberellin in tomato fruit set. J Exp Bot 60:1523–1532. doi: 10.1093/jxb/erp094

de Tournefort JP (1694) Élemens de botanique, ou méthode pour connoître les plantes. de l'Imprimerie royale, Paris

Dempewolf H, Hodgins KA, Rummell SE, et al (2012) Reproductive isolation during domestication. Plant Cell 24:2710–2717. doi: 10.1105/tpc.112.100115

DePristo MA, Banks E, Poplin R, et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491

Diamond J (2002) Evolution, consequences and future of plant and animal domestication. Nature 418:700

Doebley J (1989a) Isozymic evidence and the evolution of crop plants. In: Soltis DE, Soltis PS, Dudley TR (eds) Isozymes in Plant Biology. Springer Netherlands, Dordrecht, pp 165–191

Doebley J (1989b) Molecular evidence for a missing wild relative of maize and the introgression of its chloroplast genome into *zea perennis*. Evolution (N Y) 43:1555–1559. doi: 10.1111/j.1558-5646.1989.tb02603.x

Doebley J, Lukens L (1998) Transcriptional regulators and the evolution of plant Form. Plant Cell 10:1075. doi: 10.2307/3870712

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. Cell 127:1309–1321. doi: 10.1016/j.cell.2006.12.006

Doganlar S, Frary A, Daunay MC, et al (2002a) Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant. Genetics 161:1713–1726. doi: 12196413

Doganlar S, Frary A, Daunay MC, et al (2002b) A comparative genetic linkage map of eggplant (*Solanum melongena*) and its implications for genome evolution in the Solanaceae. Genetics 161:1697–1711

Doganlar S, Frary A, Tanksley SD (2000) The genetic basis of seed-weight variation: tomato as a model system. TAG Theor Appl Genet 100:1267–1273. doi: 10.1007/s001220051433

Duangjit J, Causse M, Sauvage C (2016) Efficiency of genomic selection for tomato fruit quality. Mol Breed 36:29. doi: 10.1007/s11032-016-0453-3

Dubin MJ, Zhang P, Meng D, et al (2015) DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife 4:e05255. doi: 10.7554/eLife.05255

Duncan NA, Pearsall DM, Benfer RA (2009) Gourd and squash artifacts yield starch grains of feasting foods from preceramic Peru. Proc Natl Acad Sci 106:13202–13206. doi: 10.1073/pnas.0903322106

Dvornyk V, Sirviö A, Mikkonen M, Savolainen O (2002) Low nucleotide diversity at the pal1 locus in the widely distributed *Pinus sylvestris*. Mol Biol Evol 19:179–188. doi: 10.1093/oxfordjournals.molbev.a004070

E. Z. Kochieva NNR (2003) Molecular Aflp analysis of the genotypes of pepper. Russ J Genet 39:1345–1348

Ellegren H, Galtier N (2016) Determinants of genetic diversity. Nat Rev Genet 17:422–433. doi: 10.1038/nrg.2016.58

Ervynck A, Dobney K, Hongo H, Meadow R (2001) Born free ? New evidence for the status of 'sus scrofa' at neolithic *Çayönü Tepesi* (Southeastern Anatolia, Turkey). Paléorient 27:47–73

Eshbaugh WH (1983) The genus *Capsicum* in Africa. Bothalia 14:845–848.

Eshbaugh WH (1993) Peppers: history and exploitation of a serendipitous new crop discovery, new crops. Wiley, New York

Eshed Y, Zamir D (1995) An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics 141:1147 LP-1162

Excoffier L, Dupanloup I, Huerta-Sánchez E, et al (2013) Robust Demographic Inference from Genomic and SNP Data. PLoS Genet 9:e1003905. doi: 10.1371/journal.pgen.1003905

Eyre-Walker A, Gaut RL, Hilton H, et al (1998) Investigation of the bottleneck leading to the domestication of maize. Proc Natl Acad Sci 95:4441–4446. doi: 10.1073/pnas.95.8.4441

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–87

Farris JS (1982) Outgroups and Parsimony. Syst Biol 31:328–334. doi: 10.1093/sysbio/31.3.328

Fernandez-Pozo N, Menda N, Edwards JD, et al (2015) The Sol Genomics Network (SGN)--from genotype to phenotype to breeding. Nucleic Acids Res 43:D1036-41. doi: 10.1093/nar/gku1195

Fernandez-Pozo N, Zheng Y, Snyder SI, et al (2017) The Tomato Expression Atlas. Bioinformatics 33:1–3. doi: 10.1093/bioinformatics/btx190

Fernie AR, Schauer N (2009) Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet 25:39–48. doi: 10.1016/j.tig.2008.10.010

Fernie AR, Tadmor Y, Zamir D (2006) Natural genetic variation for improving crop quality. Curr Opin Plant Biol 9:196–202. doi: 10.1016/j.pbi.2006.01.010

Fery RL, Thies JA (1997) Evaluation of *Capsicum chinense* Jacq. cultigens for resistance to the southern root-knot nematode. Hortscience 32:923–926

Fingerhuth KA (1832) Monographia generis Capsici. Arnz & Comp.

Finkers R, van Heusden AW, Meijer-Dekens F, et al (2007) The construction of a *Solanum habrochaites* LYC4 introgression line population and the identification of QTLs for resistance to *Botrytis cinerea*. Theor Appl Genet 114:1071–1080. doi: 10.1007/s00122-006-0500-2

Finn RD, Coggill P, Eberhardt RY, et al (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279–D285. doi: 10.1093/nar/gkv1344

Firn RD, Jones CG (2003) Natural products - a simple model to explain chemical diversity. Nat Prod Rep 20:382. doi: 10.1039/b208815k

Florez-Rueda AM, Paris M, Schmidt A, et al (2016) Genomic imprinting in the endosperm is systematically perturbed in abortive hybrid tomato seeds. Mol Biol Evol 33:2935–2946. doi: 10.1093/molbev/msw175

Foolad MR (2007) Genome mapping and molecular breeding of tomato. Int J Plant Genomics 2007:1–52. doi: 10.1155/2007/64358

Frary A (2000) fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289:85–88. doi: 10.1126/science.289.5476.85

Frary A, Doganlar S, Daunay MC (2007) 9 Eggplant. Mol Breed 5:

Frary A, Doganlar S, Daunay MC, Tanksley SD (2003) QTL analysis of morphological traits in eggplant and implications for conservation of gene function during evolution of solanaceous species. Theor Appl Genet 107:359–370. doi: 10.1007/s00122-003-1257-5

Frary A, Nesbitt C, Frary A, et al (2000) Cloning, Transgenic Expression and Function of fw2.2: a Quantitative trait locus key to the evolution of tomato fruit. Science 289:85–88

Fromm J (1997) Hormonal physiology of wood growth in willow (*Salix viminalis* L.): effects of spermine and abscisic acid. Wood Sci Technol 31:119–130. doi: 10.1007/BF00705927

Fulgione A, Koornneef M, Roux F, et al (2018) Madeiran Arabidopsis thaliana reveals ancient long-range colonization and clarifies demography in eurasia. Mol Biol Evol 35:564–574. doi: 10.1093/molbev/msx300

Fuller DQ, Allaby R (2009) Seed Dispersal and Crop Domestication: Shattering, Germination and Seasonality in Evolution under Cultivation

Fuller DQ, Denham T, Arroyo-Kalin M, et al (2014) Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record. Proc Natl Acad Sci U S A 111:6147–52. doi: 10.1073/pnas.1308937110

Furini A, Wunder J (2004) Analysis of eggplant (*Solanum melongena*)-related germplasm: Morphological and AFLP data contribute to phylogenetic interpretations and germplasm utilization. Theor Appl Genet 108:197–208. doi: 10.1007/s00122-003-1439-1

Gallusci P, Dai Z, Génard M, et al (2017) Epigenetics for plant improvement: current knowledge and modeling avenues. Trends Plant Sci. 22:610–623

Galtier N, Depaulis F, Barton NH (2000) Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 155:981–987

Gaut BS, Seymour DK, Liu Q, Zhou Y (2018) Demography and its effects on genomic variation in crop domestication. Nat Plants 4:512–520. doi: 10.1038/s41477-018-0210-1

Gayral P, Melo-Ferreira J, Glémin S, et al (2013) Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. PLoS Genet 9:e1003457. doi: 10.1371/journal.pgen.1003457

Gepts P (2004) Selection Experiment. Plant Breed 24:1–44. doi: 0-471-46892-4

Gerbault P, Allaby RG, Boivin N, et al (2014) Storytelling and story testing in domestication. Proc Natl Acad Sci 111:6159–6164. doi: 10.1073/pnas.1400425111

Gershenzon J, Dudareva N (2007) The function of terpene natural products in the natural world. Nat Chem Biol 3:408–414. doi: 10.1038/nchembio.2007.5

Gill SS, Tuteja N (2010) Polyamines and abiotic stress tolerance in plants. Plant Signal Behav 5:26–33. doi: 10.4161/psb.5.1.10291

Giuliano, The Eggplant Genome Consortium G (2017) The eggplant genome reveals paleopolyploid origin of fruit ripening. In: XIV Solanaceae and 3rd Cucurbitaceae Joint Conference - Solcuc2017. Valencia 3-6 September, pp 5–6

Glémin S, Bataillon T (2009) A comparative view of the evolution of grasses under domestication: Tansley review. New Phytol 183:273–290. doi: 10.1111/j.1469-8137.2009.02884.x

Goddard ME, Hayes BJ (2007) Genomic selection. J Anim Breed Genet 124:323–330. doi: 10.1111/j.1439-0388.2007.00702.x

Goerg GM (2011) Lambert W random variables—a new family of generalized skewed distributions with applications to risk estimation. Ann Appl Stat 5:2197–2230. doi: 10.1214/11-AOAS457

González-Pérez S, Garcés-Claver A, Mallor C, et al (2014) New insights into *Capsicum spp* relatedness and the diversification process of *Capsicum annuum* in Spain. PLoS One 9:1–23. doi: 10.1371/journal.pone.0116276

Grandillo S, Cammareri M (2016) Molecular mapping of quantitative trait loci in tomato. pp 39–73

Grandillo S, Chetelat R, Knapp S, et al (2011) *Solanum* sect. *Lycopersicon*. In: Wild Crop Relatives: Genomic and Breeding Resources. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 129–215

Grandillo S, Ku HM, Tanksley SD (1999) Identifying the loci responsible for natural variation in fruit size and shape in tomato. Theor Appl Genet 99:978–987. doi: 10.1007/s001220051405

Grandillo S, Tanksley SD (1996) QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium*. Theor Appl Genet 92:935–951. doi: 10.1007/BF00224033

Greco M, Chiappetta A, Bruno L, Bitonti MB (2012) Roles and regulation of cytokinins in tomato fruit development. J Exp Bot 63:695–709. doi: 10.1093/jxb/ers207

Grimm D, Greshake B, Kleeberger S, et al (2012) easyGWAS: An integrated interspecies platform for performing genome-wide association studies. arXiv Prepr 1–22

Grobman A, Bonavia D, Dillehay TD, et al (2012) Preceramic maize from Paredones and Huaca Prieta, Peru. Proc Natl Acad Sci 109:1755–1759. doi: 10.1073/pnas.1120270109

Gruissem W (1989) Chloroplast gene expression: how plants turn their plastids on. Cell 56:161–70. doi: 10.1016/0092-8674(89)90889-1

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695. doi: 10.1371/journal.pgen.1000695

Guzmán FA, Ayala H, Azurdia C, et al (2005) AFLP assessment of genetic diversity of genetic resources in Guatemala. Crop Sci 45:363. doi: 10.2135/cropsci2005.0363
Haak DC, Kostyun JL, Moyle LC (2014) Merging ecology and genomics to dissect diversity in wild tomatoes and their relatives. pp 273–298
Hammer K (1984) Das Domestikationssyndrom. Die Kult 32:11–34. doi: 10.1007/BF02098682
Harborne JB (1999) Recent advances in chemical ecology. Nat Prod Rep 16:509–523
Harlan JR (1992) Origins and processes of domestication. Grass Evol Domest 159:175
Harrigan GG, Martino-Catt S, Glenn KC (2007) Metabolomics, metabolic diversity and genetic variation in crops. Metabolomics 3:259–272. doi: 10.1007/s11306-007-0076-0
Haudry A, Cenci A, Ravel C, et al (2007) Grinding up wheat: A massive loss of nucleotide diversity since domestication. Mol Biol Evol 24:1506–1517. doi: 10.1093/molbev/msm077
Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443. doi: 10.3168/jds.2008-1646
Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic gelection for crop improvement. Crop Sci 49:1. doi: 10.2135/cropsci2008.08.0512
Heiser CB, Pickersgill B (1969) Names for the Cultivated *Capsicum Species* (Solanaceae). Taxon 18:277. doi: 10.2307/1218828
Heiser CB, Smith PG (1953) The cultivated C*apsicum* peppers. Econ Bot 7:214–227. doi: 10.1007/BF02984948
Henry RJ (2012) Next-generation sequencing for understanding and accelerating crop domestication. Brief Funct Genomics 11:51–56. doi: 10.1093/bfgp/elr032
Henry RJ, Nevo E (2014) Exploring natural selection to guide breeding for agriculture. Plant Biotechnol J 12:655–662. doi: 10.1111/pbi.12215
Hernández-Verdugo S, Luna-Reyes R, Oyama K (2001) Genetic structure and differentiation of wild and domesticated populations of *Capsicum annuum* (Solanaceae) from Mexico. Plant Syst Evol 226:129–142. doi: 10.1007/s006060170061
Heun M, Schäfer-Pregl R, Klawan D, et al (1997) Site of Einkorn Wheat Domestication Identified by DNA Fingerprinting. Science 278:1312–1314. doi: 10.1126/science.278.5341.1312
Hill TA, Ashrafi H, Reyes-Chin-Wo S, et al (2013) Characterization of *Capsicum annuum* genetic diversity and population structure based on parallel polymorphism discovery with a 30K unigene pepper GeneChip. PLoS One 8:56200. doi: 10.1371/journal.pone.0056200
Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231. doi: 10.1007/BF01245622
Hindorff LA, Sethupathy P, Junkins HA, et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci 106:9362–9367. doi: 10.1073/pnas.0903103106
Hirakawa H, Shirasawa K, Miyatake K, et al (2014) Draft genome sequence of eggplant (*Solanum melongena* L.): the representative solanum species indigenous to the old world. DNA Res 21:649–60. doi: 10.1093/dnares/dsu027
Hohenlohe P a., Phillips PC, Cresko W a. (2011) Using population genomics to detect selection in natural populations: Key concepts and methodological considerations. Int J Plant Sci 171:1059–1071. doi: 10.1086/656306.USING
Hollister JD, Smith LM, Guo Y-L, et al (2011) Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc Natl Acad Sci
Hufford MB, Lubinksy P, Pyhäjärvi T, et al (2013) The genomic signature of crop-wild introgression in maize. PLoS Genet 9:e1003477. doi: 10.1371/journal.pgen.1003477
Hufford MB, Martínez-Meyer E, Gaut BS, et al (2012) Inferences from the historical distribution of wild and domesticated maize provide ecological and evolutionary insight. PLoS One 7:. doi: 10.1371/journal.pone.0047659
Hulse-Kemp AM, Maheshwari S, Stoffel K, et al (2018) Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library. Hortic Res 5:4. doi: 10.1038/s41438-017-0011-0

Hung H-Y, Shannon LM, Tian F, et al (2012) ZmCCT and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. Proc Natl Acad Sci 109:E1913–E1921. doi: 10.1073/pnas.1203189109

Huxley SJ (1963) Eugenics in Evolutionary Perspective. Perspect Biol Med 6:155–187. doi: 10.1353/pbm.1963.0015

Ibiza VP, Blanca J, Cañizares J, Nuez F (2012) Taxonomy and genetic diversity of domesticated *Capsicum* species in the Andean region. Genet Resour Crop Evol 59:1077–1088. doi: 10.1007/s10722-011-9744-z

Ibiza VP, Canizares J, Nuez F (2010) EcoTILLING in *Capsicum* species: searching for new virus resistances. BMC Genomics 11:631. doi: 10.1186/1471-2164-11-631

Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci 101:10667–10672. doi: 10.1073/pnas.0401720101

Isshiki S, Suzuki S, Yamashita KI (2003) RFLP analysis of mitochondrial DNA in eggplant and related Solanum species. Genet Resour Crop Evol 50:133–137. doi: 10.1023/A:1022954229295

Itkin M, Heinig U, Tzfadia O, et al (2013) Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. Science 341:175–179. doi: 10.1126/science.1240230

Jenkins JA (1948) The origin of the cultivated tomato. Econ Bot 2:379–392. doi: 10.1007/BF02859492

Jensen R, McLeod M, Eshbaugh W, Guttman S (1979) Numerical taxonomic analyses of allozymic variation in Capsicum (Solanaceae). Taxon 28:315–327. doi: 10.2307/1219739

Jones P, Binns D, Chang HY, et al (2014) InterProScan 5: Genome-scale protein function classification. Bioinformatics 30:1236–1240. doi: 10.1093/bioinformatics/btu031

Jouganous J, Long W, Ragsdale AP, Gravel S (2017) Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. Genetics 206:1549–1567. doi: 10.1534/genetics.117.200493

Kamm JA, Terhorst J, Song YS (2017) Efficient Computation of the Joint Sample Frequency Spectra for Multiple Populations. J Comput Graph Stat 26:182–194. doi: 10.1080/10618600.2016.1159212

Karihaloo JL, Gottlieb LD (1995) Allozyme variation in the eggplant, *Solanum melongena* L. (Solanaceae). Theor Appl Genet 90:578–583. doi: 10.1007/BF00222006

Kaufman T (1994) The native languages of Mesoamerica. In: Moseley C, Asher RE, Darkes G (eds) Atlas of the world's languages. Routledge, London, UK., pp 34 – 41

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. Ecol Lett 7:1225–1241. doi: 10.1111/j.1461-0248.2004.00684.x

Kerr EA, Bailey DL (1964) Resistance to C*ladosporium fulvum* cke. obtained from wild species of tomato. Can J Bot 42:1541–1554. doi: 10.1139/b64-153

Kim S, Park M, Yeom S-I, et al (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat Genet 46:270–278. doi: 10.1038/ng.2877

Kim Y, Neilsen R (2004) Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513–1524. doi: 10.1534/genetics.103.025387

Kimura M (1983) The neutral theory of molecular evolution. Cambridge Univ Press Cambridge, Engl 367

Kirkpatrick M, Ravigné V (2002) Speciation by natural and sexual selection: models and experiments. Am Nat 159:S22–S35. doi: 10.1086/338370

Klee HJ (2010) Improving the flavor of fresh fruits: genomics, biochemistry, and biotechnology. New Phytol 187:44–56. doi: 10.1111/j.1469-8137.2010.03281.x

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, Jiménez-Gómez JM, Kimura S, et al (2013) Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. Proc Natl Acad Sci 110:E2655–E2662. doi: 10.1073/pnas.1309606110

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29. doi: 10.1186/1746-4811-9-29

Kortstee AJ, Appeldoorn NJG, Oortwijn MEP, Visser RGF (2007) Differences in regulation of carbohydrate metabolism during early fruit development between domesticated tomato and two wild relatives. Planta 226:929–939. doi: 10.1007/s00425-007-0539-6

Kraft KH, Brown CH, Nabhan GP, et al (2014) Multiple lines of evidence for the origin of domesticated chili pepper, *Capsicum annuum*, in Mexico. Proc Natl Acad Sci 111:6165–6170. doi: 10.1073/pnas.1308933111

Kudo T, Kobayashi M, Terashima S, et al (2017) TOMATOMICS: A web database for integrated omics information in tomato. Plant Cell Physiol 58:e8. doi: 10.1093/pcp/pcw207

Kurlovich BS, Rep'ev SI, Petrova MV, et al (2000) The significance of Vavilov's scientific expeditions and ideas for development and use of legume genetic resources. Plant Genet Resour Newsl 124:23–32

Labate JA, Grandillo S, Fulton T, et al (2007) Vegetables. Springer Berlin Heidelberg, Berlin, Heidelberg

Labate JA, Robertson LD, Baldo AM (2009) Multilocus sequence data reveal extensive departures from equilibrium in domesticated tomato (*Solanum lycopersicum* L.). Heredity (Edinb) 103:257–267. doi: 10.1038/hdy.2009.58

Labate JA, Robertson LD, Strickler SR, Mueller LA (2014) Genetic structure of the four wild tomato species in the *Solanum peruvianum* s.l. species complex. Genome 57:169–80. doi: 10.1139/gen-2014-0003

Ladizinsky G (1987) Pulse domestication before cultivation. Econ Bot 41:60–65

Lagriffol J, Monnier M (1985) Effects of endosperm and placenta on development of *Capsella* embryos in ovules cultivated in vitro. J Plant Physiol 118:127–137. doi: 10.1016/S0176-1617(85)80141-3

Lai J, Li R, Xu X, et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42:1027

Lam HM, Xu X, Liu X, et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–1059. doi: 10.1038/ng.715

Langenheim JH (1994) Higher plant terpenoids: A phytocentric overview of their ecological roles. J Chem Ecol 20:1223–80. doi: 10.1007/BF02059809

Langford AN (1937) The parasitism of *cladosphorium fulvum* cooke and the genetics of resistance to it. Can J Res 15c:108–128. doi: 10.1139/cjr37c-008

Langridge P, Fleury D (2011) Making the most of "omics" for crop breeding. Trends Biotechnol 29:33–40. doi: 10.1016/j.tibtech.2010.09.006

Lanteri S, Barchi L, Toppino L, et al (2014) An eggplant (*Solanum melongena* L.) high quality genome draft. In: Congress VISCEA on Applied Vegetable Genomics. Wien 19-20 February

Larson G (2014) The modern view of domestication special feature. Proc Natl Acad Sci 111:6139–6146

Laura P, Nelly D, E. HB, et al (2014) Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. Plant Biotechnol J 13:565–577. doi: 10.1111/pbi.12282

Lechner M, Findeiß S, Steiner L, et al (2011) Proteinortho : detection of ( co- ) orthologs in large-scale analysis. BMC Bioinformatics 12:124. doi: 10.1186/1471-2105-12-124

Lefebvre V, Palloix A, Rives M (1993) Nuclear RFLP between pepper cultivars (*Capsicum annuum* L.). Euphytica 71:189–199. doi: 10.1007/BF00040408

Lemmon ZH, Bukowski R, Sun Q, Doebley JF (2014) The role of cis regulatory evolution in maize domestication. PLoS Genet 10:e1004745. doi: 10.1371/journal.pgen.1004745

Lentz DL, Beaudry-Corbett MP, Aguilar MLR de, Kaplan L (1996) Foodstuffs, forests, fields, and shelter: a paleoethnobotanical analysis of vessel contents from the Ceren Site, El Salvador. Lat Am Antiq 7:247–262. doi: 10.2307/971577

Lestari P, Lee G, Ham T-H, et al (2011) Single nucleotide polymorphisms and haplotype diversity in rice sucrose synthase 3. J Hered 102:735–746. doi: 10.1093/jhered/esr094

Lester RN (1986) Taxonomy of Scarlet Eggplant, *Solanum aethiopicum* L. Acta Hortic. 182:125–132

Lester RN, Hasan SMZ (1990) The Distinction between *Solanum incanum* L. and *Solanum insanum* L. (Solanaceae). Source: Taxon 39:521–523. doi: 10.2307/1223119

Levin RA, Myers NR, Bohs L (2006) Phylogenetic relationships among the "spiny solanums" ( *Solanum* subgenus *Leptostemonum* , Solanaceae). Am J Bot 93:157–169. doi: 10.3732/ajb.93.1.157

Lewinsohn E, Gijzen M (2009) Phytochemical diversity: The sounds of silent metabolism. Plant Sci 176:161–169. doi: 10.1016/j.plantsci.2008.09.018

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175–195. doi: 10.1186/1475-925X-13-94

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993. doi: 10.1093/bioinformatics/btr509

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493

Li W, Chetelat RT (2010) A pollen factor linking inter- and intraspecific pollen rejection in t tomato. Science 330:1827–1830. doi: 10.1126/science.1197908

Li W, Chetelat RT (2015) Unilateral incompatibility gene ui1.1 encodes an S-locus F-box protein expressed in pollen of *Solanum* species. Proc Natl Acad Sci 112:4417–4422. doi: 10.1073/pnas.1423301112

Li X, Jian Y, Xie C, et al (2017) Fast diffusion of domesticated maize to temperate zones. Sci Rep 7:2077. doi: 10.1038/s41598-017-02125-0

Lin T, Zhu G, Zhang J, et al (2014) Genomic analyses provide insights into the history of tomato breeding. Nat Genet 46:1220–1226. doi: 10.1038/ng.3117

Lin Z, Li X, Shannon LM, et al (2012) Parallel domestication of the Shattering1 genes in cereals. Nat Genet 44:720

Lindblad-Toh K, Wade CM, Mikkelsen TS, et al (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819. doi: 10.1038/nature04338

Linnaeus C (1753) Species plantarum. Impensis G. C. Nauk, Stockholm

Lippman Z, Tanksley SD (2001) Dissecting the genetic pathway to extreme fruit size in tomato using a cross between the small-fruited wild species Lycopersicon pimpinellifolium and L. esculentum var. giant heirloom. Genetics 158:413 LP-422

Lippman ZB, Semel Y, Zamir D (2007) An integrated view of quantitative trait variation using tomato interspecific introgression lines. Curr Opin Genet Dev 17:545–552. doi: 10.1016/j.gde.2007.07.007

Liu J, Van Eck J, Cong B, Tanksley SD (2002) A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. Proc Natl Acad Sci 99:13302 LP-13306

Liu M, Sugimoto K, Uygun S, et al (2018) Regulatory divergence in wound-responsive gene expression between domesticated and wild tomato. Plant Cell 23:1–43. doi: 10.1105/tpc.18.00194

Liu X, Fu Y-X (2015) Exploring population size changes using SNP frequency spectra. Nat Genet 47:555

Loaiza-Figueroa F, Ritland K, Cancino JAL, Tanksley SD (1989) Patterns of genetic variation of the genus *Capsicum* (Solanaceae) in Mexico. Plant Syst Evol 165:159–188. doi: 10.1007/BF00936000

Lohmueller KE (2014) The impact of population demography and selection on the genetic architecture of complex traits. PLoS Genet 10:e1004379. doi: 10.1371/journal.pgen.1004379

Lorant A, Pedersen S, Holst I, et al (2017) The potential role of genetic assimilation during maize domestication. PLoS One 12:0–9. doi: 10.1371/journal.pone.0184202

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. doi: 10.1186/s13059-014-0550-8

Lu J, Tang T, Tang H, et al (2006) The accumulation of deleterious mutations in rice genomes: A hypothesis on the cost of domestication. Trends Genet 22:126–131. doi: 10.1016/j.tig.2006.01.004

Luckwill LC (1943) The genus Lycopersicon: an historical, biological, and taxonomic survey of the wild and cultivated tomatoes

Luikart G (1998) Distortion of allele frequency distributions provides a test for recent population bottlenecks. J Hered 89:238–247. doi: 10.1093/jhered/89.3.238

Lynch M (2010) Evolution of the mutation rate. Trends Genet 26:345–352. doi: 10.1016/j.tig.2010.05.003

Mace ES, Lester RN, Gebhardt CG (1999) AFLP analysis of genetic relationships among the cultivated eggplant, *Solanum melongena* L., and wild relatives (Solanaceae). Theor Appl Genet 99:626–633. doi: 10.1007/s001220051277

Mapelli S (1981) Changes in cytokinin in the fruits of parthenocarpic and normal tomatoes. Plant Sci Lett 22:227–233. doi: 10.1016/0304-4211(81)90235-2

Marin J-M, Pudlo P, Robert CP, Ryder RJ (2012) Approximate Bayesian computational methods. Stat Comput 22:1167–1180. doi: 10.1007/s11222-011-9288-2

Marshall DJ, Uller T (2007) When is a maternal effect adaptive? Oikos 116:1957–1963. doi: 10.1111/j.2007.0030-1299.16203.x

Marshall FB, Dobney K, Denham T, Capriles JM (2014) Evaluating the roles of directed breeding and gene flow in animal domestication. Proc Natl Acad Sci 111:6153–6158. doi: 10.1073/pnas.1312984110

Marshall JA, Knapp S, Davey MR, et al (2001) Molecular systematics of Solanum section Lycopersicum (Lycopersicon) using the nuclear ITS rDNA region. Theor Appl Genet 103:1216–1222. doi: 10.1007/s001220100671

Martin GB, Brommonschenkel SH, Chunwongse J, et al (1993) Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science 262:1432 LP-1436

Martínez-Castillo J, Zizumbo-Villarreal D, Gepts P, Colunga-GarcíaMarín P (2007) Gene Flow and Genetic Structure in the Wild–Weedy–Domesticated Complex of Phaseolus lunatus L. in its Mesoamerican Center of Domestication and Diversity. Crop Sci 47:58. doi: 10.2135/cropsci2006.04.0241

Matsuoka Y, Vigouroux Y, Goodman MM, et al (2002) A single domestication for maize shown by multilocus microsatellite genotyping. Proc Natl Acad Sci 99:6080–6084. doi: 10.1073/pnas.052125199

McCarthy DJ, Humburg P, Kanapin A, et al (2014) Choice of transcripts and software has a large effect on variant annotation. Genome Med 6:26. doi: 10.1186/gm543

McClean PE, Hanson MR (1986) A community-based annotation framework for linking solanaceae genomes with phenomes. Genetics 112:649–667

McLaren W, Gil L, Hunt SE, et al (2016) The Ensembl Variant Effect Predictor. Genome Biol 17:13534–13544. doi: 10.1186/s13059-016-0974-4

Menda N, Buels RM, Tecle I, Mueller LA (2008) A community-based annotation framework for linking solanaceae genomes with phenomes. Plant Physiol 147:1788–1799. doi: 10.1104/pp.108.119560

Mendel G (1866) Versuche über Pflanzenhybriden. Verhandlungen des naturforschenden Vereines Brunn 4:3–44

Meyer RS, Choi JY, Sanches M, et al (2016) Domestication history and geographic adaptation inferred from a SNP map of African rice. Submitt to Nat Genet 1–9. doi: 10.1038/ng.3633

Meyer RS, Duval AE, Jensen HR (2012a) Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytol 196:29–48. doi: 10.1111/j.1469-8137.2012.04253.x

Meyer RS, Karol KG, Little DP, et al (2012b) Phylogeographic relationships among Asian eggplants and new perspectives on eggplant domestication. Mol Phylogenet Evol 63:685–701. doi: 10.1016/j.ympev.2012.02.006

Meyer RS, Purugganan MD (2013) Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet 14:840–52. doi: 10.1038/nrg3605

Meyer RS, Whitaker BD, Little DP, et al (2015) Parallel reductions in phenolic constituents resulting from the domestication of eggplant. Phytochemistry 115:194–206. doi: 10.1016/j.phytochem.2015.02.006

Miller AJ, Gross BL (2011) From forest to field: Perennial fruit crop domestication. Am J Bot 98:1389–1414. doi: 10.3732/ajb.1000522

Miller J, Miller P (1768) The gardeners dictionary. Published by John and Francis Rivington as well as 23 others, London

Miller JC, Tanksley SD (1990) RFLP analysis of phylogenetic relationships and genetic variation in the genus Lycopersicon. Theor Appl Genet 80:437–48. doi: 10.1007/BF00226743

Molina J, Sikora M, Garud N, et al (2011) Molecular evidence for a single evolutionary origin of domesticated rice. Proc Natl Acad Sci U S A 108:8351–6. doi: 10.1073/pnas.1104686108

Monforte AJ, Tanksley SD (2000) Development of a set of near isogenic and backcross recombinant inbred lines containing most of the *Lycopersicon hirsutum* genome in a *L. esculentum* genetic background: A tool for gene mapping and gene discovery. Genome 43:803–813. doi: 10.1139/g00-043

Moniz De Sá M, Drouin G (1996) Phylogeny and substitution rates of angiosperm actin genes. Mol Biol Evol 13:1198–1212. doi: 10.1093/oxfordjournals.molbev.a025685

Morrell PL, Buckler ES, Ross-Ibarra J (2011) Crop genomics: advances and applications. Nat Rev Genet 13:85–96. doi: 10.1038/nrg3097

Mousavi-Derazmahalleh M, Bayer PE, Nevado B, et al (2018) Exploring the genetic and adaptive diversity of a pan-Mediterranean crop wild relative: narrow-leafed lupin. Theor Appl Genet 131:887–901. doi: 10.1007/s00122-017-3045-7

Moyers BT, Morrell PL, McKay JK (2018) Genetic costs of domestication and improvement. J Hered 109:103–116. doi: 10.1093/jhered/esx069

Moyle LC (2008) Ecological and evolutionary genomics in the wild tomatoes (*Solanum* sect. *Lycopersicon*). Evolution (N Y) 62:2995–3013. doi: 10.1111/j.1558-5646.2008.00487.x

Moyle LC (2007) Comparative genetics of potential prezygotic and postzygotic isolating barriers in a *Lycopersicon* species cross. J Hered 98:123–135. doi: 10.1093/jhered/esl062

Mueller LA (2005) The SOL Genomics Network. A comparative resource for solanaceae biology and beyond. Plant Physiol 138:1310–1317. doi: 10.1104/pp.105.060707

Mueller LA, Tanskley SD, Giovannoni JJ, et al (2005a) The tomato sequencing project, the first cornerstone of the International Solanaceae project (SOL). Comp Funct Genomics 6:153–158. doi: 10.1002/cfg.468

Mueller UG, Gerardo NM, Aanen DK, et al (2005b) The evolution of agriculture in insects. Annu Rev Ecol Evol Syst 36:563–595. doi: 10.1146/annurev.ecolsys.36.102003.152626

Muir CD, Moyle LC (2009) Antagonistic epistasis for ecophysiological trait differences between Solanum species. New Phytol 183:789–802. doi: 10.1111/j.1469-8137.2009.02949.x

Muller CH (1942) Notes on the American flora, chiefly Mexican. Am Midl Nat 27:470. doi: 10.2307/2421014

Müller CH (1940) A revision of the genus Lycopersicon. U.S. Dept. of Agriculture

Müller NA, Zhang L, Koornneef M, Jiménez-gómez JM (2018) Mutations in EID1 and LNK2 caused light-conditional clock deceleration during tomato domestication. Proc Natl Acad Sci U S A. doi: 10.1073/pnas.1801862115

Muñeoz-Falcón JE, Prohens J, Vilanova S, Nuez F (2009) Diversity in commercial varieties and landraces of black eggplants and implications for broadening the breeders' gene pool. Ann Appl Biol 154:453–465. doi: 10.1111/j.1744-7348.2009.00314.x

Muños S, Ranc N, Botton E, et al (2011) Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near WUSCHEL. Plant Physiol 156:2244–2254. doi: 10.1104/pp.111.173997

Musseau C, Just D, Jorly J, et al (2017) Identification of two new mechanisms that regulate fruit growth by cell expansion in tomato. Front Plant Sci 8:. doi: 10.3389/fpls.2017.00988

Mutegi E, Snow AA, Rajkumar M, et al (2015) Genetic diversity and population structure of wild/weedy eggplant (Solanum insanum, Solanaceae) in southern India: Implications for conservation. Am J Bot 102:140–148. doi: 10.3732/ajb.1400403

Mutlu N, Boyaci FH, Göçmen M, Abak K (2008) Development of SRAP, SRAP-RGA, RAPD and SCAR markers linked with a *Fusarium wilt* resistance gene in eggplant. Theor Appl Genet 117:1303–1312. doi: 10.1007/s00122-008-0864-6

Nabholz B, Sarah G, Sabot F, et al (2014) Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice ( *Oryza glaberrima* ). Mol Ecol 23:2210–2227. doi: 10.1111/mec.12738

Nakazato T, Bogonovich M, Moyle LC (2008) Environmental factors predict adaptive phenotypic differentiation within and between two wild andean tomatoes. Evolution (N Y) 62:774–792. doi: 10.1111/j.1558-5646.2008.00332.x

Nakazato T, Franklin RA, Kirk BC, Housworth EA (2012) Population structure, demographic history, and evolutionary patterns of a green-fruited tomato, *Solanum peruvianum* (Solanaceae), revealed by spatial genetics analyses. Am J Bot 99:1207–1216. doi: 10.3732/ajb.1100210

Nakazato T, Housworth EA (2011) Spatial genetics of wild tomato species reveals roles of the Andean geography on demographic history. Am J Bot 98:88–98. doi: 10.3732/ajb.1000272

Nakazato T, Warren DL, Moyle LC (2010) Ecological and geographic modes of species divergence in wild tomatoes. Am J Bot 97:680–693. doi: 10.3732/ajb.0900216

Nicolaï M, Cantet M, Lefebvre V, et al (2013) Genotyping a large collection of pepper (*Capsicum spp*.) with SSR loci brings new evidence for the wild origin of cultivated *C. annuum* and the structuring of genetic diversity by human selection of cultivar types. Genet Resour Crop Evol 60:2375–2390. doi: 10.1007/s10722-013-0006-0

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218. doi: 10.1146/annurev.genet.39.073003.112420

Nunome T, Ishiguro K, Yoshida T, Hirai M (2001) Mapping of fruit shape and color development traits in eggplant (*Solanum melongena* L.) based on RAPD and AFLP markers. Breed Sci 51:19–26. doi: 10.1270/jsbbs.51.19

Nunome T, Negoro S, Kono I, et al (2009) Development of SSR markers derived from SSR-enriched genomic library of eggplant (*Solanum melongena* L.). Theor Appl Genet 119:1143–1153. doi: 10.1007/s00122-009-1116-0

Nunome T, Suwabe K, Iketani H, Hirai M (2003) Identification and characterization of microsatellites in eggplant. Plant Breed 122:256–262. doi: 10.1046/j.1439-0523.2003.00816.x

O'Connor TP (1997) Working at relationships: another look at animal domestication. Antiquity 71:149–156. doi: DOI: 10.1017/S0003598X00084635

Ofner I, Lashbrooke J, Pleban T, et al (2016) *Solanum pennellii* backcross inbred lines (BILs) link small genomic bins with tomato traits. Plant J 87:151–160. doi: 10.1111/tpj.13194

Ofori DA, Gyau A, Dawson IK, et al (2014) Developing more productive African agroforestry systems and improving food and nutritional security through tree domestication. Curr. Opin. Environ. Sustain. 6:123–127

Ohmori T, Murata M, Motoyoshi F (1998) Characterization of disease resistance gene-like sequences in near-isogenic lines of tomato. Theor Appl Genet 96:331–338. doi: 10.1007/s001220050745

Ohmori T, Murata M, Motoyoshi F (1995) Identification of RAPD markers linked to the Tm-2 locus in tomato. Theor Appl Genet 307–311

Olmstead RG, Palmer JD (1997) Implications for the phylogeny, classification, and biogeography of solanaceae from cpDNA site variation. Syst Bot 22:19–29

Olmstead RG, Sweere JA, Spangler RE, et al (1999) Phylogeny and provisional classification of the Solanaceae based on chloroplast DNA. Solanaceae IV Adv Biol Util 111–137

Olsen KM, Wendel JF (2013) A Bountiful Harvest: Genomic Insights into Crop Domestication Phenotypes. Annu Rev Plant Biol 64:47–70. doi: 10.1146/annurev-arplant-050312-120048

Orozco-Cardenas ML, Narvaez-Vasquez J, Ryan CA (2001) Hydrogen peroxide acts as a second messenger for the induction of defense genes in tomato plants in response to wounding, systemin, and methyl jasmonate. Plant Cell 13:179. doi: 10.2307/3871162

Ostrander EA, Wayne RK, Freedman AH, Davis BW (2017) Demographic history, selection and functional diversity of the canine genome. Nat Rev Genet. doi: 10.1038/nrg.2017.67

Palmer JD, Zamir D (1982) Chloroplast DNA evolution and phylogenetic relationships in Lycopersicon. Proc Natl Acad Sci U S A 79:5006–5010

Palmgren MG, Edenbrandt AK, Vedel SE, et al (2015) Are we ready for back-to-nature crop breeding? Trends Plant Sci 20:155–164. doi: 10.1016/j.tplants.2014.11.003

Paran I, Aftergoot E, Shifriss C (1998) Variation in *Capsicum annuum* revealed by RAPD and AFLP markers. Euphytica 99:167–173. doi: 10.1023/A:1018301215945

Paran I, Van Der Knaap E (2007) Genetic and molecular regulation of fruit and plant domestication traits in tomato and pepper. J Exp Bot 58:3841–3852. doi: 10.1093/jxb/erm257

Paterson AH, Damon S, Hewitt JD, et al (1991) Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. Genetics 127:181 LP-197

Paterson AH, Lander ES, Hewitt JD, et al (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335:721

Patil G, Do T, Vuong TD, et al (2016) Genomic-assisted haplotype analysis and the development of high-throughput SNP markers for salinity tolerance in soybean. Sci Rep 6:19199

Pease JB, Haak DC, Hahn MW, Moyle LC (2016) Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol 14:e1002379. doi: 10.1371/journal.pbio.1002379

Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L (2013) On the accumulation of deleterious mutations during range expansions. Mol Ecol 22:5972–5982. doi: 10.1111/mec.12524

Peralta IE, Spooner DM (2001) Granule-bound starch synthase (GBSSI) gene phylogeny of wild tomatoes (*Solanum* L. section *Lycopersicun* [Mill.] Wettst. subsection *Lycopersicon*). Am J Bot 88:1888–1902

Peralta IE, Spooner DM (2000) Classification of wild tomatoes: a review. Kurtziana 28:45–54

Peralta IE, Spooner DM, Knapp S (2008) Taxonomy of wild tomatoes and their relatives (*Solanum* sect. *Lycopersicoides*, sect. *Juglandifolia*, sect. *Lycopersicon*; Solanaceae). Amer. Society of Plant Taxonomists

Perez-Fons L, Wells T, Corol DI, et al (2014) A genome-wide metabolomic resource for tomato fruit from *Solanum pennellii*. Sci Rep 4:1–8. doi: 10.1038/srep03859

Perry L, Dickau R, Zarrillo S, et al (2007) Starch fossils and the domestication and dispersal of chili peppers (*Capsicum spp*. L.) in the Americas. Science 315:986–8. doi: 10.1126/science.1136914

Pertuzé RA, Ji Y, Chetelat RT (2002) Comparative linkage map of the *Solanum lycopersicoides* and *S. sitiens* genomes and their differentiation from tomato. Genome 45:1003–1012. doi: 10.1139/g02-066

Peters SA, Bargsten JW, Szinay D, et al (2012) Structural homology in the Solanaceae: analysis of genomic regions in support of synteny studies in tomato, potato and pepper. Plant J 71:602–614. doi: 10.1111/j.1365-313X.2012.05012.x

Pichersky E, Gershenzon J (2002) The formation and function of plant volatiles: perfumes for pollinator attraction and defense. Curr. Opin. Plant Biol. 5:237–243

Pickersgill B (2007) Domestication of plants in the Americas: Insights from Mendelian and molecular genetics. Ann Bot 100:925–940. doi: 10.1093/aob/mcm193

Pickersgill B (1971) Relationships between weedy and cultivated forms in some species of chili peppers (genus *Capsicum*). Evolution (N Y) 25:683–691. doi: 10.1111/j.1558-5646.1971.tb01926.x

Pickersgill B (1997) Genetic resources and breeding of *Capsicum spp*. Euphytica 96:129–133. doi: 10.1023/A:1002913228101

Pickersgill B, Heiser CBJ (1977) Origins and distribution of plants domesticated in the New World tropics. In: Origins of agriculture. pp 803–835

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

Pimentel D, Wilson C, McCullum C, et al (1997) Economic and Environmental Benefits of Biodiversity. Bioscience 47:747–757. doi: 10.2307/1313097

Piperno DR, Ranere AJ, Holst I, et al (2009) Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. Proc Natl Acad Sci 106:5019–5024. doi: 10.1073/pnas.0812525106

Pnueli L, Carmel-Goren L, Hareven D, et al (1998) The SELF-PRUNING gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of CEN and TFL1. Development 125:1979–1989. doi: 10.1126/science.250.4983.959

Pohl MED, Piperno DR, Pope KO, Jones JG (2007) Microfossil evidence for pre-Columbian maize dispersals in the neotropics from San Andres, Tabasco, Mexico. Proc Natl Acad Sci 104:6870–6875. doi: 10.1073/pnas.0701425104

Polston JE, Cohen L, Sherwood TA, et al (2006) *Capsicum* species: symptomless hosts and reservoirs of tomato yellow leaf curl virus. Phytopathology 96:447–452. doi: 10.1094/PHYTO-96-0447

Prince JP, Pochard E, Tanksley SD (1993) Construction of a molecular linkage map of pepper and a comparison of synteny with tomato. Genome 36:404–417. doi: 10.1139/g93-056

Purcell S, Neale B, Todd-Brown K, et al (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575. doi: 10.1086/519795

Purugganan MD, Fuller DQ (2011) Archaeological data reveal slow rates of evolution during plant domestication. Evolution (N Y) 65:171–183. doi: 10.1111/j.1558-5646.2010.01093.x

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

Qi J, Liu X, Shen D, et al (2013) A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat Genet 45:1510–1515. doi: 10.1038/ng.2801

Qi X, An H, Ragsdale AP, et al (2017) Genomic inferences of domestication events are corroborated by written records in *Brassica rapa*. Mol Ecol 26:3373–3388. doi: 10.1111/mec.14131

Qin C, Yu C, Shen Y, et al (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. Proc Natl Acad Sci U S A 111:5135–5140. doi: 10.1073/pnas.1400975111

R Core Team (R Foundation for Statistical Computing) (2016) R: a Language and Environment for Statistical Computing

Rakyan VK, Down TA, Balding DJ, Beck S (2011) Epigenome-wide association studies for common human diseases. Nat Rev Genet 12:529

Ramsey J, Bradshaw HD, Schemske DW (2003) Components of reproductive isolation between the monkeyflowers *Mimulus lewisii* and *M. cardinalis* (Phrymaceae). Evolution 57:1520–34

Ranc N, Muños S, Santoni S, Causse M (2008) A clarified position for Solanum lycopersicum var. cerasiforme in the evolutionary history of tomatoes (solanaceae). BMC Plant Biol 8:130. doi: 10.1186/1471-2229-8-130

Ranc N, Muños S, Xu J, et al (2012) Genome-Wide Association Mapping in Tomato ( *Solanum lycopersicum* ) Is Possible Using Genome Admixture of Solanum lycopersicum var. cerasiforme. G3 2:853–864. doi: 10.1534/g3.112.002667

Ranere AJ, Piperno DR, Holst I, et al (2009) The cultural and chronological context of early Holocene maize and squash domestication in the Central Balsas River Valley, Mexico. Proc Natl Acad Sci 106:5014–5018. doi: 10.1073/pnas.0812590106

Ranil RHG, Prohens J, Aubriot X, et al (2016) *Solanum insanum* L. (subgenus *Leptostemonum* Bitter, Solanaceae), the neglected wild progenitor of eggplant (*S. melongena* L.): a review of taxonomy, characteristics and uses aimed at its enhancement for improved eggplant breeding. Genet Resour Crop Evol. doi: 10.1007/s10722-016-0467-z

Ranwez V, Serra A, Pot D, Chantret N (2017) Domestication reduces alternative splicing expression variations in sorghum. PLoS One 12:1–20. doi: 10.1371/journal.pone.0183454

Rapp RA, Haigler CH, Flagel L, et al (2010) Gene expression in developing fibres of Upland cotton (*Gossypium hirsutum* L.) was massively altered by domestication. BMC Biol 8:1–15. doi: 10.1186/1741-7007-8-139

Ray J (1673) Observations topographical, moral and physiological, made in a journey through part of low-countries, Germany, Italy, and France. London

Reddy AC, Venkat S, Singh TH, et al (2015) Isolation, characterization and evolution of NBS-LRR encoding disease-resistance gene analogs in eggplant against bacterial wilt. Eur J Plant Pathol 143:417–426. doi: 10.1007/s10658-015-0693-9

Rendón-Anaya M, Herrera-Estrella A (2018) The advantage of parallel selection of domestication genes to accelerate crop improvement. Genome Biol 19:18–20. doi: 10.1186/s13059-018-1537-7

Rick CM (1960) Hybridization between *lycopersicon esculentum* and *solanum pennellii*: phylogenetic and cytogenetic significance. Proc Natl Acad Sci 46:78–82. doi: 10.1073/pnas.46.1.78

Rick CM (1990) Perspectives from plant genetics: The tomato genetics stock center. Genet Resour Risk 11–19

Rick CM (1988) Tomato-like nightshades: affinities, autoecology, and breeders' opportunities. Econ Bot 42:145–154

Rick CM, Chetelat RT (1995) Utilization of related wild species for tomato improvement. Acta Hortic 21–38. doi: 10.17660/ActaHortic.1995.412.1

Rick CM, Fobes JF (1975) Allozyme variation in the cultivated tomato and closely related species. Bull Torrey Bot Club 102:376. doi: 10.2307/2484764

Rick CM, Tanksley SD (1981) Genetic variation in *Solanum pennellii*: comparisons with two other sympatric tomato species. Plant Syst Evol 139:11–45. doi: 10.1007/BF00983920

Rinaldi R, Van Deynze A, Portis E, et al (2016) New insights on eggplant/tomato/pepper synteny and identification of eggplant and pepper orthologous QTL. Front Plant Sci 7:. doi: 10.3389/fpls.2016.01031

Rindos D (1983) The origins of agriculture: an evolutionary perspective. Academic Press, Inc., London

Robbins MD, Sim S-C, Yang W, et al (2011) Mapping and linkage disequilibrium analysis with a genome-wide collection of SNPs that detect polymorphism in cultivated tomato. J Exp Bot 62:1831–1845. doi: 10.1093/jxb/erq367

Rodriguez F, Wu F, Ané C, et al (2009) Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? BMC Evol Biol 9:191. doi: 10.1186/1471-2148-9-191

Rodriguez GR, Munos S, Anderson C, et al (2011) Distribution of *SUN*, *OVATE*, *LC* and *FAS* in the Tomato Germplasm and the Relationship to Fruit Shape Diversity. Plant Physiol 156:275–285. doi: 10.1104/pp.110.167577

Rose JKC, Bennett AB (1999) Cooperative disassembly of the cellulose-xyloglucan network of plant cell walls: Parallels between cell expansion and fruit ripening. Trends Plant Sci 4:176–183. doi: 10.1016/S1360-1385(99)01405-3

Roselius K, Stephan W, Städler T (2005) The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. Genetics 171:753–763. doi: 10.1534/genetics.105.043877

Ross-Ibarra J, Morrell PL, Gaut BS (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc Natl Acad Sci U S A 104:8641–8648. doi: 10.1073/pnas.0700643104

Roth MM (2017) Variability of hybrid seed failure in wild tomatoes (*Solanum* sect. *Lycopersicon*): phenotypic and molecular signatures in the developing endosperm. PhD thesis 24694, ETH Zurich, Switzerland

Rousseaux MC, Jones CM, Adams D, et al (2005) QTL analysis of fruit antioxidants in tomato using Lycopersicon pennellii introgression lines. Theor Appl Genet 111:1396–1408. doi: 10.1007/s00122-005-0071-7

Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the species barrier: Genomic hotspots of introgression between two highly divergent ciona intestinalis species. Mol Biol Evol 30:1574–1587. doi: 10.1093/molbev/mst066

Rowley-Conwy P, Layton R (2011) Foraging and farming as niche construction: Stable and unstable adaptations. Philos Trans R Soc B Biol Sci 366:849–862. doi: 10.1098/rstb.2010.0307

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, et al (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol 34:3299–3302. doi: 10.1093/molbev/msx248

Ruggieri V, Francese G, Sacco A, et al (2014) An association mapping approach to identify favourable alleles for tomato fruit quality breeding. BMC Plant Biol 14:337. doi: 10.1186/s12870-014-0337-9

Sabeti PC, Varilly P, Fry B, et al (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–8. doi: 10.1038/nature06250

Sacco A, Ruggieri V, Parisi M, et al (2015) Exploring a tomato landraces collection for fruit-related traits by the aid of a high-throughput genomic platform. PLoS One 10:e0137139. doi: 10.1371/journal.pone.0137139

Sahu KK, Chattopadhyay D (2017) Genome-wide sequence variations between wild and cultivated tomato species revisited by whole genome sequence mapping. BMC Genomics 18:430. doi: 10.1186/s12864-017-3822-3

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sakata Y, Lester RN (1997) Chloroplast DNA diversity in brinjal eggplant (*Solanum melongena* L.) and related species. Euphytica 97:295–301. doi: 10.1023/A:1003000612441

Sakata Y, Nishio T, Matthews PJ (1991) Chloroplast DNA analysis of eggplant (*Solanum melongena*) and related species for their taxonomic affinity. Euphytica 55:21–26. doi: 10.1016/0341-8162(88)90005-7

Saliba-Colombani V, Causse M, Langlois D, et al (2001) Genetic analysis of organoleptic quality in fresh market tomato. 1. Mapping QTLs for physical and chemical traits. TAG Theor Appl Genet 102:259–272. doi: 10.1007/s001220051643

Sang T (2011) Toward the Domestication of Lignocellulosic Energy Crops: Learning from Food Crop Domestication. J. Integr. Plant Biol. 53:96–104

Sauvage C, Rau A, Aichholz C, et al (2017) Domestication rewired gene expression and nucleotide diversity patterns in tomato. Plant J 91:631–645. doi: 10.1111/tpj.13592

Sauvage C, Segura V, Bauchet G, et al (2014) Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. Plant Physiol 165:1120–1132. doi: 10.1104/pp.114.241521
Savolainen O, Pyhäjärvi T, Knürr T (2007) Gene Flow and Local Adaptation in Trees. Annu Rev Ecol Evol Syst 38:595–619. doi: 10.1146/annurev.ecolsys.38.091206.095646
Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. Nat Genet 46:919. doi: 10.1038/ng.3015
Schmidt MH-W, Vogel A, Denton AK, et al (2017) De novo assembly of a new Solanum pennellii accession using nanopore sequencing. Plant Cell 29:2336–2348. doi: 10.1105/tpc.17.00521
Schraiber JG, Akey JM (2015) Methods and models for unravelling human evolutionary history. Nat Rev Genet 16:727
Seah S, Yaghoobi J, Rossi M, et al (2004) The nematode-resistance gene, Mi-1, is associated with an inverted chromosomal segment in susceptible compared to resistant tomato. Theor Appl Genet 108:1635–1642. doi: 10.1007/s00122-004-1594-z
Sedivy EJ, Wu F, Hanzawa Y (2017) Soybean domestication: the origin, genetic architecture and molecular bases. New Phytol 214:539–553. doi: 10.1111/nph.14418
Seehausen O, Butlin RK, Keller I, et al (2014) Genomics and the origin of species. Nat Rev Genet 15:176
Shani E, Ben-Gera H, Shleizer-Burko S, et al (2010) Cytokinin regulates compound leaf development in tomato. Plant Cell 22:3206–3217. doi: 10.1105/tpc.110.078253
Shirasawa K, Fukuoka H, Matsunaga H, et al (2013) Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato. DNA Res 20:593–603. doi: 10.1093/dnares/dst033
Shrestha R, Gómez-Ariza J, Brambilla V, Fornara F (2014) Molecular control of seasonal flowering in rice, arabidopsis and temperate cereals. Ann Bot 114:1445–1458
Sim S-CC, Robbins MD, Van Deynze A, et al (2011) Population structure and genetic differentiation associated with breeding history and selection in tomato (*Solanum lycopersicum* L.). Heredity (Edinb) 106:927–935. doi: 10.1038/hdy.2010.139
Sim SC, Durstewitz G, Plieske J, et al (2012) Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS One 7:. doi: 10.1371/journal.pone.0040563
Simons AJ, Leakey RRB (2004) Tree domestication in tropical agroforestry. In: Agroforestry Systems. Springer, Dordrecht, pp 167–181
Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35. doi: 10.1017/S0016672300014634
Smith NJH, Williams JT, Plucknett DL, Talbot JP (1992) Tropical forests and their crops. Comstock Pub. Associates
Smith O, Nicholson W V, Kistler L, et al (2018) A domestication history of dynamic adaptation and genomic deterioration in sorghum. bioRxiv 336503. doi: 10.1101/336503
Smith PG, Heiser CB (1957) Taxonomy of *Capsicum sinense* Jacq. and the geographic distribution of the cultivated *Capsicum* species. Bull Torrey Bot Club 84:413. doi: 10.2307/2482971
Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 38:1409–1438
Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: Modelling gene flow. Nat Rev Genet 14:404–414. doi: 10.1038/nrg3446
Soyk S, Lemmon ZH, Oved M, et al (2017) Bypassing Negative Epistasis on Yield in Tomato Imposed by a Domestication Gene. Cell 1–14. doi: 10.1016/j.cell.2017.04.032
Spooner D, Peralta IE, Knapp S (2005) Comparisonof AFLPs with other markers for phylogenetic inference in wild tomatoes. Taxon 54:43–61. doi: 10.2307/25065301
Spooner DM, Anderson GJ, Jansen RK (1993) Chloroplast DNA evidence for the interrelationships of tomtoes, potatoes, and pepinos (Solanaceae). Am J Bot 80:676. doi: 10.2307/2445438
Stadler T (2008) Lineages-through-time plots of neutral models for speciation. Math Biosci 216:163–171. doi: 10.1016/j.mbs.2008.09.006

Stetter MG, Gates DJ, Mei W, Ross-Ibarra J (2017) How to make a domesticate. Curr Biol 27:R896–R900. doi: 10.1016/j.cub.2017.06.048

Stevens MA, Rick CM (1986) Genetics and breeding. In: The Tomato Crop. Springer Netherlands, Dordrecht, pp 35–109

Stevens R, Buret M, Duffé P, et al (2007) Candidate genes and quantitative trait loci affecting fruit ascorbic acid content in three tomato populations. Plant Physiol 143:1943 LP-1953

Suśruta, Bhishagratna K (1907) An English translation of the Sushruta samhita, based on original Sanskrit text: Edited and published by Kaviraj Kunja Lal Bhishagratna. With a full and comprehensive introd., translation of different readings, notes, comperative views, index, glossary an. Calcutta

Swanson-Wagner R, Briskine R, Schaefer R, et al (2012) Reshaping of the maize transcriptome by domestication. Proc Natl Acad Sci 109:11878–11883. doi: 10.1073/pnas.1201961109

Swinnen G, Goossens A, Pauwels L (2016) Lessons from domestication: targeting cis-regulatory elements for crop improvement. Trends Plant Sci 21:506–515. doi: 10.1016/j.tplants.2016.01.014

Syfert MM, Castañeda-Álvarez NP, Khoury CK, et al (2016) Crop wild relatives of the brinjal eggplant (*Solanum melongena*): Poorly represented in genebanks and many species at risk of extinction. Am J Bot 103:635–651. doi: 10.3732/ajb.1500539

Tadmor Y, Fridman E, Gur A, et al (2002) Identification of malodorous, a wild species allele affecting tomato aroma that was selected against during domestication. J Agric Food Chem 50:2005–2009. doi: 10.1021/jf011237x

Tajima F (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics 143:1457–1465

Tam SM, Causse M, Garchery C, et al (2007) The distribution of copia-type retrotransposons and the evolutionary history of tomato and related wild species. J Evol Biol 20:1056–1072. doi: 10.1111/j.1420-9101.2007.01293.x

Tam SM, Mhiri C, Vogelaar A, et al (2005) Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. Theor Appl Genet 110:819–831. doi: 10.1007/s00122-004-1837-z

Tanksley SD (2004) The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. Plant Cell 16:S181–S189. doi: 10.1105/tpc.018119

Tanksley SD, Ganal MW, Prince JP, et al (1992) High density molecular linkage maps of the tomato and potato genomes. Genetics 132:1141 LP-1160

Tecle IY, Menda N, Buels RM, et al (2010) solQTL: a tool for QTL analysis, visualization and linking to genomes at SGN database. BMC Bioinformatics 11:525. doi: 10.1186/1471-2105-11-525

Tedja MS, Wojciechowski R, Hysi PG, et al (2018) Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error. Nat Genet 50:834–848. doi: 10.1038/s41588-018-0127-7

Tellier A, Laurent SJY, Lainer H, et al (2011) Inference of seed bank parameters in two wild tomato species using ecological and genetic data. Proc Natl Acad Sci 108:17052–17057. doi: 10.1073/pnas.1111266108

Terhorst J, Kamm JA, Song YS (2017) Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet 49:303–309. doi: 10.1038/ng.3748

The Computational Pan-Genomics Consortium ., Marschall T, Marz M, et al (2016) Computational pan-genomics: status, promises and challenges. Brief Bioinform bbw089. doi: 10.1093/bib/bbw089

The Tomato Genome Consortium, Sato S, Tabata S, et al (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641. doi: 10.1038/nature11119

Tieman D, Zhu G, Resende MFRR, et al (2017) A chemical genetic roadmap to improved tomato flavor. Science 355:391–394. doi: 10.1126/science.aal1556

Van der Auwera GA, Carneiro MO, Hartl C, et al (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. In: Current Protocols in Bioinformatics. John Wiley & Sons, Inc., Hoboken, NJ, USA, p 11.10.1-11.10.33

van der Knaap E, Lippman ZB, Tanksley SD (2002) Extremely elongated tomato fruit controlled by four quantitative trait loci with epistatic interactions. Theor Appl Genet 104:241–247. doi: 10.1007/s00122-001-0776-1

Van Inghelandt D, Melchinger AE, Martinant J-P, Stich B (2012) Genome-wide association mapping of flowering time and northern corn leaf blight (*Setosphaeria turcica*) resistance in a vast commercial maize germplasm set. BMC Plant Biol 12:56. doi: 10.1186/1471-2229-12-56

Van Zonneveld M, Ramirez M, Williams DE, et al (2015) Screening genetic resources of *Capsicum* peppers in their primary center of diversity in Bolivia and Peru. PLoS One 10:1–23. doi: 10.1371/journal.pone.0134663

Vavilov NI (1926) Studies on the origin of cultivated plants. In: Bulletin of applied Botany and plant-Breeding VOL XVI, part. 2. Leningrad, p 248

Verbelen JP, Vissenberg K, Kerstens S, Le JIE (2001) Cell expansion in the epidermis: microtubules, cellulose orientation and wall loosening enzymes. J Plant Physiol 158:537–543. doi: 10.1078/0176-1617-00277

Verlaan MG, Hutton SF, Ibrahem RM, et al (2013) The tomato yellow leaf curl virus resistance genes Ty-1 and Ty-3 are allelic and code for DFDGD-Class RNA-Dependent RNA Polymerases. PLoS Genet 9:. doi: 10.1371/journal.pgen.1003399

Videvall E, Sletvold N, Hagenblad J, et al (2016) Strong maternal effects on gene expression in arabidopsis lyrata hybrids. Mol Biol Evol 33:984–994. doi: 10.1093/molbev/msv342

Vigne JD (2011) The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. Comptes Rendus - Biol 334:171–181. doi: 10.1016/j.crvi.2010.12.009

Vincent H, Wiersema J, Kell S, et al (2013) A prioritized crop wild relative inventory to help underpin global food security. Biol Conserv 167:265–275. doi: 10.1016/j.biocon.2013.08.011

Víquez-Zamora M, Vosman B, van de Geest H, et al (2013) Tomato breeding in the genomics era: insights from a SNP array. BMC Genomics 14:354. doi: 10.1186/1471-2164-14-354

Vorontsova M, Knapp S (2012) A new species of *Solanum* (Solanaceae) from South Africa related to the cultivated eggplant. PhytoKeys 8:1. doi: 10.3897/phytokeys.8.2462

Wade CM, Giulotto E, Sigurdsson S, et al (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326:865–867

Walsh BM, Hoot SB (2001) Phylogenetic relationships of *Capsicum* (Solanaceae) using DNA sequences from two noncoding regions: the chloroplast atpB - rbcL spacer region and nuclear waxy introns. Int J Plant Sci 162:1409–1418. doi: 10.1086/323273

Wang M, Li W, Fang C, et al (2018) Parallel selection on a dormancy gene during domestication of crops from multiple families. Nat Genet 50:1435–1441. doi: 10.1038/s41588-018-0229-2

Wang Y, Diehl A, Wu F, et al (2008) Sequencing and comparative analysis of a conserved syntenic segment in the solanaceae. Genetics 180:391–408. doi: 10.1534/genetics.108.087981

Wang Z, Cao H, Sun Y, et al (2013) Arabidopsis paired amphipathic helix proteins SNL1 and SNL2 redundantly regulate primary seed dormancy via abscisic acid-ethylene antagonism mediated by histone deacetylation. Plant Cell 25:149–66. doi: 10.1105/tpc.112.108191

Weese TL, Bohs L (2010) Eggplant origins: out of Africa, into the Orient. Taxon 59:49–56. doi: 10.2307/27757050

Welch AJ, Wiley AE, James HF, et al (2012) Ancient DNA reveals genetic stability despite demographic decline: 3,000 years of population history in the endemic Hawaiian petrel. Mol Biol Evol 29:3729–3740. doi: 10.1093/molbev/mss185

Winter D, Vinegar B, Nahal H, et al (2007) An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS One 2:e718. doi: 10.1371/journal.pone.0000718

Wright SI (2005) The effects of artificial selection on the maize genome. Science 308:1310–1314. doi: 10.1126/science.1107891

Wu F, Eannetta NT, Xu Y, et al (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*. Theor Appl Genet 118:1279–1293. doi: 10.1007/s00122-009-0980-y

Wu F, Mueller LA, Crouzillat D, et al (2006) Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: A test case in the euasterid plant clade. Genetics 174:1407–1420. doi: 10.1534/genetics.106.062455

Wu J, Wang Y, Xu J, et al (2018) Diversification and independent domestication of Asian and European pears. Genome Biol 19:77. doi: 10.1186/s13059-018-1452-y

Xia H, Camus-Kulandaivelu L, Stephan W, et al (2010) Nucleotide diversity patterns of local adaptation at drought-related candidate genes in wild tomatoes. Mol Ecol 19:4144–4154. doi: 10.1111/j.1365-294X.2010.04762.x

Xiao H, Jiang N, Schaffner E, et al (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science 319:1527 LP-1530

Xing S, Tao C, Song Z, et al (2018) Coexpression network revealing the plasticity and robustness of population transcriptome during the initial stage of domesticating energy crop *Miscanthus lutarioriparius*. Plant Mol Biol 97:489–506. doi: 10.1007/s11103-018-0754-5

Xu X, Liu X, Ge S, et al (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30:105–111. doi: 10.1038/nbt.2050

Yamaguchi M (1983) World vegetables : principles, production and nutritive values. 704

Yamamoto E, Matsunaga H, Onogi A, et al (2017) Efficiency of genomic selection for breeding population design and phenotype prediction in tomato. Heredity (Edinb) 118:202–209. doi: 10.1038/hdy.2016.84

Yamamoto E, Matsunaga H, Onogi A, et al (2016) A simulation-based breeding design that uses whole-genome prediction in tomato. Sci Rep 6:19454. doi: 10.1038/srep19454

Ye J, Wang X, Hu T, et al (2017) An InDel in the promoter of Al-ACTIVATED MALATE TRANSPORTER9 selected during tomato domestication determines fruit malate contents and aluminum tolerance. Plant Cell 29:2249–2268. doi: 10.1105/tpc.17.00211

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11:R14. doi: 10.1186/gb-2010-11-2-r14

Young ND, Tanksley SD (1989) RFLP analysis of the size of chromosomal segments retained around the Tm-2 locus of tomato during backcross breeding. Theor Appl Genet 77:353–359. doi: 10.1007/BF00305828

Zamir D (2001) Improving plant breeding with exotic genetic libraries. Nat Rev Genet 2:983

Zeder MA (2015) Core questions in domestication research. Proc Natl Acad Sci U S A 112:3191–3198. doi: 10.1073/pnas.1501711112

Zeder MA (2006) Central questions in the domestication of plants and animals. Evol Anthropol 15:105–117. doi: 10.1002/evan.20101

Zeder MA (2012) The domestication of animals. J Anthropol Res 68:161–190. doi: 10.3998/jar.0521004.0068.201

Zeggini E, Scott LJ, Saxena R, et al (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40:638–645. doi: 10.1038/ng.120

Zewdie Y, Bosland P (2000) Capsaicinoid inheritance in an interspecific hybridization of *Capsicum annuum* x *C. chinense*. J Am Soc Hortic Sci 125:448–453

Zhang J, Yang D, Li M, Shi L (2016a) Metabolic profiles reveal changes in wild and cultivated soybean seedling leaves under salt stress. PLoS One 11:. doi: 10.1371/journal.pone.0159622

Zhang J, Zhao J, Xu Y, et al (2015) Genome-wide association mapping for tomato volatiles positively contributing to tomato flavor. Front Plant Sci 6:. doi: 10.3389/fpls.2015.01042

Zhang X, Zhang Z, Gu X, et al (2016b) Genetic diversity of pepper (*Capsicum spp*.) germplasm resources in China reflects selection for cultivar types and spatial distribution. J Integr Agric 15:1991–2001. doi: 10.1016/S2095-3119(16)61364-3

Zhao Q, Feng Q, Lu H, et al (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50:278–284. doi: 10.1038/s41588-018-0041-z

Zheng X, Levine D, Shen J, et al (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28:3326–3328. doi: 10.1093/bioinformatics/bts606

Zhu G, Wang S, Huang Z, et al (2018) Rewiring of the fruit metabolome in tomato breeding. Cell 172:249–255.e12. doi: 10.1016/j.cell.2017.12.019

Zhu Q, Zheng X, Luo J, et al (2007) Multilocus Analysis of Nucleotide Variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. Mol Biol Evol 24:875–888. doi: 10.1093/molbev/msm005

Zohary D (2004) Unconscious selection and the evolution of domesticated plants unconscious selection and the evolution of domesticated plants. Econ Bot 58:5–10. doi: 10.1663/0013-0001

Zohary D (1989) Pulse domestication and cereal domestication: How different are they? Econ Bot 43:31–34. doi: 10.1007/BF02859322

Zouine M, Maza E, Djari A, et al (2017) TomExpress, a unified tomato RNA-Seq platform for visualization of expression data, clustering and correlation networks. Plant J 92:727–735. doi: 10.1111/tpj.13711

Zsögön A, Čermák T, Naves ER, et al (2018) De novo domestication of wild tomato using genome editing. Nat Biotechnol. doi: 10.1038/nbt.4272

Zuriaga E, Blanca J, Nuez F (2009) Classification and phylogenetic relationships in *Solanum* section *Lycopersicon* based on AFLP and two nuclear gene sequences. Genet Resour Crop Evol 56:663–678. doi: 10.1007/s10722-008-9392-0

# APPENDIX 1: Detailed description of the 92 accessions available for the studies

EGGPLANT

| Name | species | Country | Location | Chapter 2 | Chapter 3 |
|---|---|---|---|---|---|
| MM0014 | *S. melongena* | Greece | EU | | Crop |
| MM0498 | *S. insanum* | Japan | ASIA | | |
| MM0620 | *S. melongena* | India | ASIA | | Crop |
| MM0668 | *S. campylacanthum* | Zimbabwe | AFR. | Out | |
| MM0669 | *S. insanum* | India | ASIA | | Wild |
| MM0675 | *S. melongena* | India | ASIA | | |
| MM0686 | *S. insanum* | Indonésie | ASIA | | Wild |
| MM0693 | *S. insanum* | Sri Lanka | ASIA | | Wild |
| MM0694 | *S. melongena* | India | ASIA | | |
| MM0703 | *S. campylacanthum* | Kenya | AFR. | Out | |
| MM0709 | *S. insanum* | Malaysia | ASIA | | Wild |
| MM0710 | *S. insanum* | Thailand | ASIA | | Wild |
| MM0730 | *S. melongena* | India | ASIA | | Crop |
| MM1192 | *S. insanum* | Madagascar | AFR. | | Wild |
| MM1407 | *S. insanum* | Sri Lanka | ASIA | | Wild |
| MM1572 | *S. melongena* | Thailand | ASIA | | Crop |
| MM1592 | *S. melongena* | India | ASIA | | |
| MM1678 | *S. insanum* | Thailand | ASIA | | Wild |
| MM1789 | *S. insanum* | Vietnam | ASIA | | Wild |
| MM1803 | *S. melongena* | Egypt | AFR. | | Crop |
| MM1826 | *S. melongena* | China | ASIA | | Crop |
| MM1831 | *S. melongena* | China | ASIA | | Crop |
| MM1838 | *S. insanum* | Vietnam | ASIA | | Wild |
| MM1900 | *S. insanum* | Thailand | ASIA | | Wild |

PEPPER

| Name | species | Country | Location | Chapter 2 | Chapter 3 |
|---|---|---|---|---|---|
| PM0076 | *C. annuum* | France | EU | | Crop |
| PM0441 | *C. baccatum* | South Ameri | AFR. | Out | |
| PM0549 | *C. annuum* | Hungary | EU | | Crop |
| PM0568 | *C. annuum* | Italy | EU | | Crop |
| PM0609 | *C. annuum* | Mexico | Am.C | | Crop |
| PM0641 | *C. annuum glabriusculum* | Copsta Rica | Am.C | Wild | |
| PM0647 | *C. annuum glabriusculum* | Mexico | Am.C | | Crop |
| PM0648 | *C. annuum glabriusculum* | USA (Florida | US | Wild | |
| PM0663 | *C. annuum glabriusculum* | Mexico | Am.C | Wild | |
| PM0669 | *C. annuum glabriusculum* | Panama | Am.C | Wild | |
| PM0702 | *C. annuum* | Mexico | Am.C | | Crop |
| PM0828 | *C. annuum glabriusculum* | NA | X | | Crop |
| PM0910 | *C. annuum* | Turquey | EU | | Crop |
| PM0952 | *C. frutescens* | Guatemala | Am.C | | Wild |
| PM1022 | *C. baccatum* | Chili | Am.S | Out | |
| PM1093 | *C. chinense* | Mexico | Am.C | | Wild |
| PM1100 | *C. annuum* | Cuba | Am.C | Crop | |
| PM1219 | *C. frutescens* | Nepal | ASIE | | Wild |
| PM1269 | *C. chacoense* | Bolivia | Am.S | Out | |
| PM1272 | *C. chacoense* | Bolivia | Am.S | Out | |
| PM1565 | *C. annuum* | China | ASIA | Crop | |
| PM1573 | *C. annuum* | Sudan | AFR. | Crop | |
| PM1600 | *C. annuum* | Mexico | Am.C | Crop | |
| PM1621 | *C. chinense* | Cameroun | AFR. | | Wild |

TOMATO

| Name | Species | Clade | data from | LA - name | Country | Location | | Chapter 3 | Chapter 2 |
|---|---|---|---|---|---|---|---|---|---|
| LACMVSel | *S. peruvianum* | peruvianum | | *LACMVSel* | ? | ? | | | Wild |
| LASS1 | *S. pimpinellifolium* | Esculentum | | LA1589 | Peru | La Libertad | | Wild | |
| LA0107 | *S. peruvianum* | peruvianum | Pease et al. | LA0107 | Peru | Lima | | | |
| LA0444 | *S. peruvianum* | peruvianum | Pease et al. | LA0444 | Peru | Ica | | | |
| LA1269 | *S. pimpinellifolium* | Esculentum | Pease et al. | LA1269 | Peru | Lima | | | |
| LA1274 | *S. peruvianum* | Peruvianum | | LA1274 | Peru | Lima | | | Wild |
| LA1283 | *S. corneliomulleri* | Peruvianum | | LA1283 | Peru | Lima | *S. corneliomulleri* | | Wild |
| LA1358 | *S. huaylasense* | Peruvianum | | LA1358 | Peru | Ancash | *S. huaylasense* | | Wild |
| LA1360 | *S. huaylasense* | Peruvianum | Pease et al. | LA1360 | Peru | Ancash | *S. huaylasense* | | |
| LA1364 | *S. huaylasense* | Peruvianum | Pease et al. | LA1364 | Peru | Ancash | *S. huaylasense* | | |
| LA1365 | *S. huaylasense* | Peruvianum | | LA1365 | Peru | Ancash | *S. huaylasense* | | Wild |
| LA1552 | *S. corneliomulleri* | Peruvianum | | LA1552 | Peru | Lima | *S. corneliomulleri* | | Wild |
| LA1969 | *S. chilense* | Peruvianum | | LA1969 | Peru | Tacna | *S. chilense* | | |
| LA2744 | *S. peruvianum* | Peruvianum | Pease et al. | LA2744 | Chile | Arica-Parinacota | | | |
| LA2933 | *S. lycopersicum* | Esculentum | Pease et al. | LA2933 | Ecuador | Manabi | | | |
| LA2964 | *S. peruvianum* | Peruvianum | Pease et al. | LA2964 | Peru | Tacna | | | |
| LA3475 | *S. lycopersicum* | Esculentum | Pease et al. | LA3475 or M-82 | Modern Culti | / | | Crop | |
| LA4117 | *S. chilense* | Peruvianum | Pease et al. | LA4117 | Chile | Antofagasta | *S. chilense* | | |
| LASC1 | *S. lycopersicum* | Esculentum | | *Levovil* | Modern Culti | / | | | Crop |
| LASC10 | *S. lycopersicum* | Esculentum | | LA0409 | Ecuador | Guayaquil | | | Crop |
| LASC2 | *S. lycopersicum* | Esculentum | | *Stupicke Polni Rane* | Modern Culti | / | | | Crop |
| LASC3 | *S. lycopersicum* | Esculentum | | *Plovdiv 24A* | Modern Culti | / | | | Crop |
| LASC4 | *S. lycopersicum* | Esculentum | | LA1420 | Ecuador | Lago Agrio | | | Crop |
| LASC5 | *S. lycopersicum* | Esculentum | | *Criollo* | Modern Culti | / | | | Crop |
| LASC6 | *S. lycopersicum* | Esculentum | | LA0147 | Honduras | Tegucigalpa | | | |
| LASC7 | *S. lycopersicum* | Esculentum | | *Cervil* | Modern Culti | / | | | |
| LASC8 | *S. lycopersicum* | Esculentum | | *FERUM* | Modern Culti | / | | | Crop |
| LASC9 | *S. lycopersicum* | Esculentum | | LA0767 | Guatemala | Quetzaltenango | | | Crop |
| LASS10 | *S. pimpinellifolium* | Esculentum | | LA1245 | Ecuador | El Oro | | Wild | |
| LASS2 | *S. pimpinellifolium* | Esculentum | | LA1478 | Peru | Piura | | Wild | |
| LASS3 | *S. pimpinellifolium* | Esculentum | | LA1582 | Peru | Lambayeque | | Wild | |
| LASS4 | *S. pimpinellifolium* | Esculentum | | LA1593 | Peru | La Libertad | | Wild | |
| LASS5 | *S. pimpinellifolium* | Esculentum | | LA1602 | Peru | Lima | | Wild | |
| LASS6 | *S. pimpinellifolium* | Esculentum | | LA1729 | Peru | Ica | | Wild | |
| LASS7 | *S. pimpinellifolium* | Esculentum | | *L.pimpi.site10(F300(* | ? | ? | | Wild | |
| LASS8 | *S. pimpinellifolium* | Esculentum | | *732292* | ? | ? | | Wild | |
| LASS9 | *S. pimpinellifolium* | Esculentum | | LA0411 | Ecuador | Los Rios | | | |
| LA0407 | *S. habrochaites* | Hirsitum | Pease et al. | LA0407 | Ecuador | Guayas | | | |
| LA0429 | *S. chesmanii* | Esculentum | Pease et al. | LA0429 | Galápagos | Is. Santa Cruz | | | |
| LA0436 | *S. galapagense* | Esculentum | Pease et al. | LA0436 | Galápagos | Is. Isabella | | | |
| LA0716 | *S. penellii* | Hirsitum | | LA0716 | Peru | Arequipa | | | |
| LA1028 | *S. chmielewskii* | Arcanum | Pease et al. | LA1028 | Peru | Apurimac | | | |
| LA1223 | *S. habrochaites* | Hirsitum | | LA1223 | Ecuador | Alausi | | | |
| LA1297 | *S. penellii* | Hirsitum | | LA1297 | Peru | Pucara | | | |
| LA1316 | *S. chmielewskii* | Arcanum | Pease et al. | LA1316 | Peru | Ayacucho | | | |
| LA1321 | *S. neorickii* | Arcanum | | LA1321 | Peru | Curahuasi | | | |
| LA1322 | *S. neorickii* | Arcanum | Pease et al. | LA1322 | Peru | Apurimac | | | |
| LA1367 | *S. penellii* | Hirsitum | | LA1367 | Peru | Santa Eulalia | | | |
| LA1401 | *S. chesmanii* | Esculentum | | LA1401 | Ecuador | Isabella | | | |
| LA1412 | *S. chesmanii* | Esculentum | | LA1412 | Ecuador | San Cristobal | | | |
| LA1447 | *S. chesmanii* | Esculentum | | LA1447 | Ecuador | Santa Cruz | | | |
| LA1777 | *S. habrochaites* | Hirsitum | | LA1777 | Peru | Rio Casma | | | |
| LA1840 | *S. chmielewskii* | Arcanum | | LA1840 | ? | ? | | | |
| LA2133 | *S. neorickii* | Arcanum | | LA2133 | Ecuador | Ona | | | |
| LA2172 | *S. arcanum* | Arcanum | Pease et al. | LA2172 | Peru | Cajamarca | | | |
| LA2325 | *S. neorickii* | Arcanum | | LA2325 | Peru | Above Balsas | | | |
| LA2548 | *S. arcanum* | Arcanum | | LA2548 | Peru | La Moyuna | | | |
| LA2680 | *S. chmielewskii* | Arcanum | | LA2680 | Peru | Apurimac | | | |
| LA2951 | *S. lycopersicoides* | outgroup | Pease et al. | LA2951 | Chile | Tarapaca | | | |
| LA3124 | *S. chesmanii* | Esculentum | Pease et al. | LA3124 | Galápagos | Is. Santa Fe | | | |
| LA3778 | *S. penellii* | Hirsitum | Pease et al. | LA3778 | Peru | Ica | | | |
| LA3863 | *S. habrochaites* | Hirsitum | | LA3863 | ? | ? | | | |
| LA3909 | *S. galapagense* | Esculentum | Pease et al. | LA3909 | Galápagos | Is. Bartolome | | | |
| LA4116 | *S. sitiens* | outgroup | Pease et al. | LA4116 | Chile | Antofagasta | | | |
| LA4126 | *S. lycopersicoides* | outgroup | Pease et al. | LA4126 | Chile | Antofagasta | | | |
| LAHirsutumB | *S. habrochaites* | Hirsitum | | LAHirsutumB | ? | ? | | | |
| LAPI247087 | *S. habrochaites* | Hirsitum | | LAPI247087 | ? | ? | | | |

# APPENDIX 2: Supplementary tables related to the chapter 2

**Table S1: Detailed data about the studied accessions. The species and the location of origins are listed in separated tables for the three accessions: a. the eggplant accessions, b. the pepper accessions and c. the tomato accessions. The accessions in white background are the crop species, the ones in grey background are the wild species and the ones in green background are the outgroups.**

| S1*a* | Name | Species | IDs | POP | Country | Extracted from paper: |
|---|---|---|---|---|---|---|
| | MM0014 | *S. melongena* | | Crop | Greece | |
| | MM0620 | *S. melongena* | | Crop | India | |
| | MM0668 | *S. campylacanthum* | | Out | Zimbabwe | |
| | MM0669 | *S. insanum* | | Wild | India | |
| | MM0686 | *S. insanum* | | Wild | Indonésie | |
| | MM0693 | *S. insanum* | | Wild | Sri Lanka | |
| | MM0703 | *S. campylacanthum* | | Out | Kenya | |
| | MM0709 | *S. insanum* | | Wild | Malaysia | |
| | MM0730 | *S. melongena* | | Crop | India | |
| | MM1192 | *S. insanum* | | Wild | Madagascar | |
| | MM1407 | *S. insanum* | | Wild | Sri Lanka | |
| | MM1572 | *S. melongena* | | Crop | Thailand | |
| | MM1803 | *S. melongena* | | Crop | Egypt | |
| | MM1826 | *S. melongena* | | Crop | China | |
| | MM1831 | *S. melongena* | | Crop | China | |

| S1*b* | Name | Species | IDs | POP | Country | Extracted from paper: |
|---|---|---|---|---|---|---|
| | PM0076 | *C. annuum* | | Crop | France | |
| | PM0441 | *C. microcarpum* | | Out | South America | |
| | PM0549 | *C. annuum* | | Crop | Hungary | |
| | PM0568 | *C. annuum* | | Crop | Italy | |
| | PM0609 | *C. annuum* | | Crop | Mexico | |
| | PM0641 | *C. annuum glab.* | | Wild | Copsta Rica | |
| | PM0648 | *C. annuum glab.* | | Wild | USA (Florida) | |
| | PM0663 | *C. annuum glab.* | | Wild | Mexico | |
| | PM0669 | *C. annuum glab.* | | Wild | Panama | |
| | PM0702 | *C. annuum* | | Crop | Mexico | |
| | PM0910 | *C. annuum* | | Crop | Turquey | |
| | PM1022 | *C. baccatum* | | Out | Chili | |
| | PM1100 | *C. annuum* | | Crop | Cuba | |
| | PM1269 | *C. chacoense* | | Out | Bolivia | |
| | PM1272 | *C. chacoense* | | Out | Bolivia | |
| | PM1565 | *C. annuum* | | Crop | China | |
| | PM1573 | *C. annuum* | | Crop | Sudan | |
| | PM1600 | *C. annuum / re-defined* | | Crop | Mexico | |

**S1*c***

| Name | Species | IDs | POP | Country | Extracted from paper: |
|---|---|---|---|---|---|
| LASS1 | *S. pimpinellifolium* | LA1589 | Wild | Peru | |
| LA3475 | *S. lycopersicum* | LA3475 or M-8 | Crop | Modern Cultivars | Pease et al. |
| LASC1 | *S. lycopersicum* | *Levovil* | Crop | Modern Cultivars | |
| LASC10 | *S. lycopersicum* | LA0409 | Crop | Ecuador | |
| LASC2 | *S. lycopersicum* | *Stupicke Polni* | Crop | Modern Cultivars | |
| LASC3 | *S. lycopersicum* | *Plovdiv 24A* | Crop | Modern Cultivars | |
| LASC4 | *S. lycopersicum* | LA1420 | Crop | Ecuador | |
| LASC5 | *S. lycopersicum* | *Criollo* | Crop | Modern Cultivars | |
| LASC8 | *S. lycopersicum* | *FERUM* | Crop | Modern Cultivars | |
| LASC9 | *S. lycopersicum* | LA0767 | Crop | Guatemala | |
| LASS10 | *S. pimpinellifolium* | LA1245 | Wild | Ecuador | Sauvage et al. |
| LASS2 | *S. pimpinellifolium* | LA1478 | Wild | Peru | |
| LASS3 | *S. pimpinellifolium* | LA1582 | Wild | Peru | |
| LASS4 | *S. pimpinellifolium* | LA1593 | Wild | Peru | |
| LASS5 | *S. pimpinellifolium* | LA1602 | Wild | Peru | |
| LASS6 | *S. pimpinellifolium* | LA1729 | Wild | Peru | |
| LASS7 | *S. pimpinellifolium* | *L.pimpi.site10(* | Wild | ? | |
| LASS8 | *S. pimpinellifolium* | *732292* | Wild | ? | |
| LA2951 | *S. lycopersicoïdes* | LA2951 | Out | Chile | |
| LA4116 | *S. sitiens* | LA4116 | Out | Chile | Pease et al. |
| LA4126 | *S. lycopersicoïdes* | LA4126 | Out | Chile | |

**Table S2: The best of the 50 runs is selected according to the log likelihood and the detailed posterior parameters are listed, for each species, for the 10 models. The models in green are the best models, and the pink values are the posteriors stumbling over boundaries due to overfitting. k is the number of parameters of the model; n is the number of finished inferences of the 50 independent runs; AIC is the Akaike Information Criterion (calculated as 2*k - 2*logL); Ne correspond to the effective population size relative to the Nref (Ne W: wild, Ne C: crop, Ne cB: crop after bottleneck, Ne cW: wild after growth/decline, Ne cE: crop after growth/decline); m correspond to the migration rate (mCW: migration rate from wild to crop, mWC: migration rate from crop to wild); T represents the times in generations relative to the Nref (Ts: duration of the first epoch from the split to next demographic event, Tb: duration of the second epoch, Te: duration of the third epoch); Theta is related to the Nref, the length of the sequences used to obtain the jAFS and to the mutation rate.**

***EGGPLANT***

| | Timing of gene flow | Population size change (crop) | | | Population size change (wild) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | | epoch 1 | epoch 2 | epoch 3 | epoch 1 | epoch 2 | epoch 3 | k | n | LogL Best | AIC Best | NeCb | NeCe | NeC | NeWe | NeW |
| SI_C | Strict Isolation | Constant | | | Constant | | | 3 | 49 | -1396,9253 | 2797,8506 | | | | | |
| IM_C | | Constant | | | Constant | | | 4 | 49 | -618,4386 | 1244,8773 | | | | | |
| IM_E | | Exp. | | | Exp. | | | 6 | 49 | -592,1706 | 1196,3412 | | | 3,09832872 | | 3,34806337 |
| IM_E_E | | Exp. | Exp. | | Exp. | Exp. | | 9 | 39 | -420,5433 | 859,0865 | | 0,23526433 | 1,02354646 | 0,11245924 | 5,72166322 |
| IM_C_E_E | Isolation w Migration | Constant | Exp. | Exp. | Constant | Exp. | Exp. | 10 | 41 | -412,9164 | 845,8328 | | 0,08668406 | 1,40860452 | 0,03402956 | 6,775008 |
| IM_BcCw | | Bott. | | | Constant | | | 5 | 50 | -618,2834 | 1246,5669 | 0,98157913 | | | | |
| IM_BcCw_E | | Bott. | Exp. | | Constant | Exp. | | 8 | 42 | -489,9206 | 995,8412 | 0,99909432 | | 3,61574092 | | 11,911988 |
| IM_C_BcCw_E | | Constant | Bott. | Exp. | Constant | Constant | Exp. | 9 | 43 | -489,7588 | 997,5176 | 0,95274507 | | 3,708581 | | 11,9825019 |
| IM2_BcCw_E | | Bott. | Exp. | | Constant | Exp. | | 10 | 36 | -360,471 | 740,9419 | 0,99846226 | | 3,94297612 | | 11,900175 |
| IM2_C_BcCw_E | | Constant | Bott. | Exp. | Constant | Constant | Exp. | 11 | 41 | -361,4321 | 744,8641 | 0,77465021 | | 4,4697764 | | 11,9716504 |

***PEPPER***

| | Timing of gene flow | Population size change (crop) | | | Population size change (wild) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | | epoch 1 | epoch 2 | epoch 3 | epoch 1 | epoch 2 | epoch 3 | k | n | LogL Best | AIC Best | NeCb | NeCe | NeC | NeWe | NeW |
| SI_C | Strict Isolation | Constant | | | Constant | | | 3 | 50 | -9324,4472 | 18652,8944 | | | | | |
| IM_C | | Constant | | | Constant | | | 4 | 50 | -7443,1359 | 14894,2718 | | | | | |
| IM_E | | Exp. | | | Exp. | | | 6 | 44 | -1694,071 | 3400,143 | | | 0,10851515 | | 1,31903338 |
| IM_E_E | | Exp. | Exp. | | Exp. | Exp. | | 9 | 47 | -1438,792 | 2895,583 | | 0,04985331 | 0,9657626 | 17,3292828 | 7,37231873 |
| IM_C_E_E | Isolation w Migration | Constant | Exp. | Exp. | Constant | Exp. | Exp. | 10 | 45 | -1172,482 | 2364,963 | | 0,02350949 | 0,3223132 | 4,25631609 | 0,54036755 |
| IM_BcCw | | Bott. | | | Constant | | | 5 | 48 | -1707,545 | 3425,09 | 0,0532525 | | | | |
| IM_BcCw_E | | Bott. | Exp. | | Constant | Exp. | | 8 | 46 | -1689,586 | 3395,171 | 0,06263031 | | 1,14459315 | | 0,13002994 |
| IM_C_BcCw_E | | Constant | Bott. | Exp. | Constant | Constant | Exp. | 9 | 48 | -1171,982 | 2361,965 | 0,06936104 | | 1,07488997 | | 0,31968567 |
| IM2_BcCw_E | | Bott. | Exp. | | Constant | Exp. | | 10 | 46 | -1409,62 | 2839,241 | 0,07274005 | | 0,17350946 | | 1,52884859 |
| IM2_C_BcCw_E | | Constant | Bott. | Exp. | Constant | Constant | Exp. | 11 | 49 | -1128,566 | 2279,131 | 0,07494818 | | 3,97161662 | | 0,09368654 |

***TOMATO***

| | Timing of gene flow | Population size change (crop) | | | Population size change (wild) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | | epoch 1 | epoch 2 | epoch 3 | epoch 1 | epoch 2 | epoch 3 | k | n | LogL Best | AIC Best | NeCb | NeCe | NeC | NeWe | NeW |
| SI_C | Strict Isolation | Constant | | | Constant | | | 3 | 50 | -7741,0291 | 15486,058 | | | | | |
| IM_C | | Constant | | | Constant | | | 4 | 50 | -5585,9486 | 11179,897 | | | | | |
| IM_E | | Exp. | | | Exp. | | | 6 | 50 | -3073,7269 | 6159,454 | | | 1,3637721 | | 6,82580166 |
| IM_E_E | | Exp. | Exp. | | Exp. | Exp. | | 9 | 44 | -2295,7205 | 4609,441 | | 0,04047524 | 8,04395626 | 6,10584102 | 5,43570422 |
| IM_C_E_E | Isolation w Migration | Constant | Exp. | Exp. | Constant | Exp. | Exp. | 10 | 40 | -2153,3 | 4326,599 | | 0,02597537 | 4,0923371 | 14,4179956 | 1,26096325 |
| IM_BcCw | | Bott. | | | Constant | | | 5 | 50 | -3371,6125 | 6753,225 | 0,25334727 | | | | |
| IM_BcCw_E | | Bott. | Exp. | | Constant | Exp. | | 8 | 43 | -2359,5783 | 4735,157 | 6,50E-05 | | 1,91224824 | | 2,22869926 |
| IM_C_BcCw_E | | Constant | Bott. | Exp. | Constant | Constant | Exp. | 9 | 44 | -2362,8207 | 4743,641 | 0,00011387 | | 2,07401793 | | 2,19833197 |
| IM2_BcCw_E | | Bott. | Exp. | | Constant | Exp. | | 10 | 38 | -2239,9959 | 4499,992 | 0,18214298 | | 2,49854658 | | 1,80028572 |
| IM2_C_BcCw_E | | Constant | Bott. | Exp. | Constant | Constant | Exp. | 11 | 43 | -2240,0005 | 4502,001 | 0,18209596 | | 2,49864615 | | 1,80013408 |

***EGGPLANT***

| Model | mCW | mWC | Ts | Tb | Te | O | theta | | |
|---|---|---|---|---|---|---|---|---|---|
| SI_C | | | 0,5561442 | | | 0,92099732 | 1679,61015 | | |
| IM_C | 0,09568515 | 0,42626317 | 3,54970646 | | | 0,92676802 | 1312,92864 | | |
| IM_E | 0,03894137 | 0,13950495 | 11,9141908 | | | 0,92261288 | 477,558966 | | |
| IM_E_E | 0,10091517 | 1,11465642 | 0,57406272 | | 0,17156028 | 0,93933401 | 2628,70073 | | |
| IM_C_E_E | 0,10358491 | 0,98631466 | 0,75262946 | 3,54E-06 | 0,2434023 | 0,9366054 | 2414,60137 | | |
| IM_BcCw | 0,09919004 | 0,42579696 | 3,49203853 | | | 0,92684175 | 1326,0203 | | |
| IM_BcCw_E | 0,09384827 | 0,29797271 | 6,37528655 | 0,24049327 | | 0,93107578 | 942,656419 | | |
| IM_C_BcCw_E | 0,09436546 | 0,29022093 | 6,21837161 | 0,0100691 | 0,24613733 | 0,9281696 | 938,777681 | mCW2 | mWC2 |
| IM2_BcCw_E | 0,08468629 | 2,09E-05 | 2,60334988 | 0,12223578 | | 0,90963613 | 1112,19118 | 0,03772029 | 1,63889878 |
| IM2_C_BcCw_E | 0,08484856 | 9,28E-19 | 2,76798334 | 0,00081525 | 0,1557003 | 0,91026372 | 1071,92773 | 0,05449807 | 1,29247249 |

***PEPPER***

| Model | mCW | mWC | Ts | Tb | Te | O | theta | | |
|---|---|---|---|---|---|---|---|---|---|
| SI_C | | | 0,44625876 | | | 0,999994 | 6135,00439 | | |
| IM_C | 0,0297585 | 0,85723965 | 1,70813549 | | | 0,99992902 | 5658,30319 | | |
| IM_E | 3,79484368 | 1,2951701 | 4,0398496 | | | 0,99990448 | 14638,1507 | | |
| IM_E_E | 0,49364584 | 0,10287841 | 1,78E-05 | | 3,41399034 | 0,99991541 | 2503,4097 | | |
| IM_C_E_E | 2,30748205 | 1,19621732 | 7,90062965 | 0,04122404 | 0,12679112 | 0,99999082 | 14914,2009 | | |
| IM_BcCw | 9,67588038 | 1,80665882 | 6,76324561 | | | 0,99989449 | 21382,5251 | | |
| IM_BcCw_E | 8,05189241 | 1,83236101 | 8,62494062 | 0,00020719 | | 0,99993982 | 21103,4069 | | |
| IM_C_BcCw_E | 2,3755401 | 1,30273607 | 8,52419059 | 0,1283919 | 0,02202778 | 0,99980947 | 17203,2587 | mCW2 | mWC2 |
| IM2_BcCw_E | 8,95797074 | 1,57883031 | 3,14970165 | 0,0919509 | | 0,99988546 | 16128,685 | 2,29445196 | 0,00047774 |
| IM2_C_BcCw_E | 2,39695707 | 1,41485594 | 4,53521509 | 0,15269807 | 0,00612828 | 0,99868221 | 18478,0371 | 8,91097978 | 3,25E-69 |

***TOMATO***

| Model | mCW | mWC | Ts | Tb | Te | O | theta | | |
|---|---|---|---|---|---|---|---|---|---|
| SI_C | | | 0,58466574 | | | 0,97824854 | 2893,58599 | | |
| IM_C | 0,14732436 | 0,24340543 | 2,60462695 | | | 0,97519507 | 2374,20392 | | |
| IM_E | 0,13179898 | 0,02869978 | 11,7108204 | | | 0,97253054 | 729,421766 | | |
| IM_E_E | 0,18710759 | 0,01928112 | 11,9058555 | | 0,97515427 | 0,97315917 | 761,950906 | | |
| IM_C_E_E | 0,3884743 | 0,05437568 | 11,9832628 | 0,89403699 | 0,43566097 | 0,97846674 | 1552,2148 | | |
| IM_BcCw | 0,73563446 | 0,12479914 | 1,3569158 | | | 0,98304435 | 3554,40014 | | |
| IM_BcCw_E | 0,59788535 | 0,06767559 | 2,22800111 | 0,72740173 | | 0,97642098 | 2470,63012 | | |
| IM_C_BcCw_E | 0,6302721 | 0,06902763 | 0,02338929 | 2,1396507 | 0,61452614 | 0,97669676 | 2553,31321 | mCW2 | mWC2 |
| IM2_BcCw_E | 2,59E-47 | 0,05104093 | 0,93387177 | 0,04099098 | | 0,97622703 | 3398,73313 | 2,20337528 | 0,32311394 |
| IM2_C_BcCw_E | 6,16E-06 | 0,05094692 | 5,18E-05 | 0,93377317 | 0,04101572 | 0,97622668 | 3398,66625 | 2,20211985 | 0,32336635 |

**Table S3: Detailed boundaries and prior probabilities are listed for each parameters for each species. All other details match Table S2.**

| | bound min | | | bound max | | | Start / prior |
|---|---|---|---|---|---|---|---|
| **Parameters** | Eggplant | Pepper | Tomato | Eggplant | Pepper | Tomato | |
| **NeC** | **12** | | | **1,00E-04** | | | **1** |
| **NeW** | **12** | | | **1,00E-04** | | | **1** |
| **NeCb** | **1** | | | 1,00E-05 | | 1,00E-06 | **0,1** |
| **Ts** | **12** | | | **0** | | | **0,5** |
| **Tb** | **12** | | | **0** | | | **0,5** |
| **Te** | **12** | | | **0** | | | **0,5** |
| **mCW** | 8 | 10 | 5 | **0** | | | **1** |
| **mWC** | 8 | 10 | 5 | **0** | | | **1** |
| **O** | **1** | | | **0** | | | **0,8** |
| **NeCe** | **20** | | | **1,00E-04** | | | **1** |
| **NeWe** | **20** | | | **1,00E-04** | | | **1** |
| **mCW2** | 10 | | 5 | **0** | | | **1** |
| **mWC2** | 10 | | 5 | **0** | | | **1** |

**Table S4: Detailed bootstraps results (x1,000) of the Godambe method, on the two best models of each species. For each parameter, the best estimate and its standard deviation obtained by bootstrap is provided. The best model is indicated in the first row, the second-best in the second row. All other details match Table S2.**

| | Model | AIC | NeCb | sd | NeCe | sd | NeC | sd | NeWe | sd | NeW | sd | mCW | sd | mWC | sd |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Eggplant*** | IM_E_E | 859,087 | | | 0,23526 | 0,02146 | 1,02355 | 0,11490 | 0,11246 | 0,00671 | 5,72166 | 1,17605 | 0,10092 | 0,01567 | 1,11466 | 0,05016 |
| | IM_C_E_E | 845,833 | | | 0,08668 | 0,01832 | 1,40860 | 0,17862 | 0,03403 | 0,00526 | 6,77501 | 0,50433 | 0,10358 | 0,01564 | 0,98631 | 0,03828 |
| ***Pepper*** | IM_C_BcCw_E | 2361,97000 | 0,06936 | 0,00817 | | | 0,06936 | 0,01194 | | | 0,31969 | 0,63761 | 2,37554 | 1,37081 | 1,30274 | 0,17065 |
| | IM_C_E_E | 2364,96000 | | | 0,02351 | 0,00303 | 0,32231 | 0,20485 | 4,25632 | *36,62907* | 0,54037 | *2,22966* | 2,30748 | 0,77964 | 1,19622 | 0,17979 |
| ***Tomato*** | IM2_BcCw_E | 4499,99 | 0,18214 | 0,70603 | | | 2,49855 | *15,10039* | | | 1,80029 | *3,57078* | 0,00000 | *3,64614* | 0,05104 | *0,37012* |
| | IM2_C_BcCw_ | 4502 | 0,18210 | 1,05062 | | | 2,49865 | *17,40429* | | | 1,80013 | *3,14756* | 0,00001 | *4,55851* | 0,05095 | *0,26270* |

| | Model | - | mCW2 | sd | mWC2 | sd | Ts | sd | Tb | sd | Te | sd | O | sd | theta | sd |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Eggplant*** | IM_E_E | - | | | | | 0,57406 | 0,04732 | | | 0,17156 | 0,01603 | 0,93933 | 0,00418 | 2628,70073 | 71,21550 |
| | IM_C_E_E | - | | | | | 0,75263 | 0,08466 | 0,00000 | 0,02374 | 0,24340 | 0,01314 | 0,93661 | 0,00411 | 2414,60137 | 64,98215 |
| ***Pepper*** | IM_C_BcCw_E | - | | | | | 8,52419 | 3,52031 | 0,12839 | 0,01753 | 0,02203 | 0,00645 | 0,99981 | 0,04378 | 17203,25873 | 6288,64601 |
| | IM_C_E_E | - | | | | | 7,90063 | 2,26767 | 0,04122 | 0,03338 | 0,12679 | *0,19675* | 0,99999 | 0,02396 | 14914,20094 | 322,34575 |
| ***Tomato*** | IM2_BcCw_E | - | 2,20338 | *11,15622* | 0,32311 | 0,35490 | 0,93387 | 0,19608 | 0,04099 | 0,09559 | | | 0,97623 | 0,04496 | 3398,73313 | 702,47329 |
| | IM2_C_BcCw_ | - | 2,20212 | *18,07361* | 0,32337 | 0,59382 | 0,00005 | *5,27406* | 0,93377 | *4,45849* | 0,04102 | 0,20964 | 0,97623 | 0,03744 | 3398,66625 | 245,60911 |

**Table S5: Mapping summary statistics on mapped, properly paired and singletons reads of all the studied accessions aligned to their reference genome, in the three species. The accessions in white background are the crop species, the ones in grey background are the wild species and the ones in green background are the outgroups.**

| *EGGPLANT* | *PEPPER* | *TOMATO* |
|---|---|---|
| **MM0014** | **PM0076** | **LA2951** |
| 31456707 + 0 mapped (80.25% : N/A) | 34351236 + 0 mapped (73.04% : N/A) | 22789761 + 0 mapped (78.21% : N/A) |
| 28100798 + 0 properly paired (72.02% : N/A) | 30761560 + 0 properly paired (65.57% : N/A) | 21365802 + 0 properly paired (73.48% : N/A) |
| 2575196 + 0 singletons (6.60% : N/A) | 2726499 + 0 singletons (5.81% : N/A) | 1053564 + 0 singletons (3.62% : N/A) |
| **MM0620** | **PM0441** | **LA3475** |
| 27511008 + 0 mapped (81.05% : N/A) | 27237790 + 0 mapped (69.46% : N/A) | 27161841 + 0 mapped (81.51% : N/A) |
| 24958758 + 0 properly paired (73.86% : N/A) | 23646072 + 0 properly paired (60.46% : N/A) | 25422756 + 0 properly paired (76.46% : N/A) |
| 1984248 + 0 singletons (5.87% : N/A) | 2710531 + 0 singletons (6.93% : N/A) | 1385123 + 0 singletons (4.17% : N/A) |
| **MM0668** | **PM0549** | **LA4116** |
| 28148339 + 0 mapped (80.27% : N/A) | 38684906 + 0 mapped (68.97% : N/A) | 23356001 + 0 mapped (71.97% : N/A) |
| 25256560 + 0 properly paired (72.35% : N/A) | 33682080 + 0 properly paired (60.19% : N/A) | 21649454 + 0 properly paired (66.83% : N/A) |
| 2233806 + 0 singletons (6.40% : N/A) | 3894904 + 0 singletons (6.96% : N/A) | 1336193 + 0 singletons (4.12% : N/A) |
| **MM0669** | **PM0568** | **LA4126** |
| 22070060 + 0 mapped (80.33% : N/A) | 22848497 + 0 mapped (73.28% : N/A) | 17367668 + 0 mapped (78.52% : N/A) |
| 19840690 + 0 properly paired (72.54% : N/A) | 20372920 + 0 properly paired (65.50% : N/A) | 16325154 + 0 properly paired (73.95% : N/A) |
| 1724537 + 0 singletons (6.30% : N/A) | 1962187 + 0 singletons (6.31% : N/A) | 786754 + 0 singletons (3.56% : N/A) |
| **MM0686** | **PM0609** | **LASC10** |
| 32088427 + 0 mapped (79.94% : N/A) | 25152847 + 0 mapped (72.97% : N/A) | 16150482 + 0 mapped (81.47% : N/A) |
| 28979926 + 0 properly paired (72.52% : N/A) | 22244024 + 0 properly paired (64.70% : N/A) | 14660212 + 0 properly paired (74.29% : N/A) |
| 2411486 + 0 singletons (6.03% : N/A) | 2265063 + 0 singletons (6.59% : N/A) | 1028254 + 0 singletons (5.21% : N/A) |
| **MM0693** | **PM0641** | **LASC1** |
| 34481776 + 0 mapped (76.66% : N/A) | 15082082 + 0 mapped (73.00% : N/A) | 23577221 + 0 mapped (84.88% : N/A) |
| 30847302 + 0 properly paired (68.86% : N/A) | 13340574 + 0 properly paired (64.72% : N/A) | 21490774 + 0 properly paired (77.72% : N/A) |
| 2779315 + 0 singletons (6.20% : N/A) | 1397850 + 0 singletons (6.78% : N/A) | 1436160 + 0 singletons (5.19% : N/A) |
| **MM0703** | **PM0648** | **LASC2** |
| 23081420 + 0 mapped (80.81% : N/A) | 21881252 + 0 mapped (69.81% : N/A) | 18644911 + 0 mapped (85.10% : N/A) |
| 20740570 + 0 properly paired (72.93% : N/A) | 19313966 + 0 properly paired (61.76% : N/A) | 17040114 + 0 properly paired (78.12% : N/A) |
| 1758606 + 0 singletons (6.18% : N/A) | 2076065 + 0 singletons (6.64% : N/A) | 1077622 + 0 singletons (4.94% : N/A) |
| **MM0709** | **PM0663** | **LASC3** |
| 24708645 + 0 mapped (80.92% : N/A) | 21198832 + 0 mapped (75.60% : N/A) | 19515571 + 0 mapped (84.22% : N/A) |
| 22230632 + 0 properly paired (73.17% : N/A) | 19134438 + 0 properly paired (68.41% : N/A) | 17820406 + 0 properly paired (77.23% : N/A) |
| 1807917 + 0 singletons (5.95% : N/A) | 1555594 + 0 singletons (5.56% : N/A) | 1170354 + 0 singletons (5.07% : N/A) |
| **MM0710** | **PM0669** | **LASC4** |
| 26753971 + 0 mapped (80.16% : N/A) | 20979962 + 0 mapped (68.88% : N/A) | 13095682 + 0 mapped (84.36% : N/A) |
| 23919582 + 0 properly paired (72.11% : N/A) | 18136590 + 0 properly paired (59.70% : N/A) | 12015946 + 0 properly paired (77.74% : N/A) |
| 1905102 + 0 singletons (5.74% : N/A) | 2202693 + 0 singletons (7.25% : N/A) | 728235 + 0 singletons (4.71% : N/A) |
| **MM0730** | **PM0702** | **LASC5** |
| 22920320 + 0 mapped (81.47% : N/A) | 25549332 + 0 mapped (71.72% : N/A) | 17897038 + 0 mapped (84.61% : N/A) |
| 20687624 + 0 properly paired (73.86% : N/A) | 22715908 + 0 properly paired (63.93% : N/A) | 16433892 + 0 properly paired (78.04% : N/A) |
| 1673126 + 0 singletons (5.97% : N/A) | 2197797 + 0 singletons (6.19% : N/A) | 951286 + 0 singletons (4.52% : N/A) |
| **MM1192** | **PM0910** | **LASC8** |
| 32686082 + 0 mapped (81.31% : N/A) | 28314720 + 0 mapped (71.97% : N/A) | 7419023 + 0 mapped (76.23% : N/A) |
| 29054800 + 0 properly paired (72.77% : N/A) | 25392088 + 0 properly paired (64.70% : N/A) | 6664180 + 0 properly paired (68.76% : N/A) |
| 2345442 + 0 singletons (5.87% : N/A) | 2293676 + 0 singletons (5.84% : N/A) | 496526 + 0 singletons (5.12% : N/A) |
| **MM1407** | **PM1022** | **LASC9** |
| 23825171 + 0 mapped (80.41% : N/A) | 27020074 + 0 mapped (73.91% : N/A) | 13514836 + 0 mapped (82.72% : N/A) |
| 21619368 + 0 properly paired (73.29% : N/A) | 23922574 + 0 properly paired (65.61% : N/A) | 12227882 + 0 properly paired (75.15% : N/A) |
| 1674289 + 0 singletons (5.68% : N/A) | 2366338 + 0 singletons (6.49% : N/A) | 926345 + 0 singletons (5.69% : N/A) |
| **MM1572** | **PM1100** | **LASS10** |
| 22589726 + 0 mapped (74.69% : N/A) | 27838757 + 0 mapped (72.72% : N/A) | 36652930 + 0 mapped (82.04% : N/A) |
| 20419022 + 0 properly paired (67.80% : N/A) | 24681030 + 0 properly paired (64.63% : N/A) | 32308550 + 0 properly paired (72.50% : N/A) |
| 1667793 + 0 singletons (5.54% : N/A) | 2475860 + 0 singletons (6.48% : N/A) | 3367547 + 0 singletons (7.56% : N/A) |

***EGGPLANT***

**MM1803**

25541703 + 0 mapped (81.54% : N/A)
23083828 + 0 properly paired (74.03% : N/A)
1868910 + 0 singletons (5.99% : N/A)

**MM1826**

29975267 + 0 mapped (81.04% : N/A)
27060626 + 0 properly paired (73.50% : N/A)
2232206 + 0 singletons (6.06% : N/A)

**MM1831**

33789446 + 0 mapped (81.55% : N/A)
30613360 + 0 properly paired (74.22% : N/A)
2417573 + 0 singletons (5.86% : N/A)

***PEPPER***

**PM1269**

24730583 + 0 mapped (73.94% : N/A)
22056866 + 0 properly paired (66.11% : N/A)
2091304 + 0 singletons (6.27% : N/A)

**PM1272**

19263819 + 0 mapped (66.83% : N/A)
16726842 + 0 properly paired (58.16% : N/A)
1999721 + 0 singletons (6.95% : N/A)

**PM1565**

25595710 + 0 mapped (73.55% : N/A)
22736268 + 0 properly paired (65.49% : N/A)
2227599 + 0 singletons (6.42% : N/A)

**PM1573**

25258354 + 0 mapped (75.19% : N/A)
22615560 + 0 properly paired (67.48% : N/A)
2054664 + 0 singletons (6.13% : N/A)

**PM1600**

20632797 + 0 mapped (73.24% : N/A)
18284918 + 0 properly paired (65.11% : N/A)
1601419 + 0 singletons (5.70% : N/A)

***TOMATO***

**LASS1**

27369583 + 0 mapped (81.77% : N/A)
24535978 + 0 properly paired (73.46% : N/A)
2213925 + 0 singletons (6.63% : N/A)

**LASS2**

23929543 + 0 mapped (81.52% : N/A)
21299560 + 0 properly paired (72.71% : N/A)
2058219 + 0 singletons (7.03% : N/A)

**LASS3**

20816204 + 0 mapped (81.33% : N/A)
18485840 + 0 properly paired (72.39% : N/A)
1764936 + 0 singletons (6.91% : N/A)

**LASS4**

27709461 + 0 mapped (83.04% : N/A)
25035364 + 0 properly paired (75.19% : N/A)
2047140 + 0 singletons (6.15% : N/A)

**LASS5**

33926248 + 0 mapped (83.34% : N/A)
30672836 + 0 properly paired (75.51% : N/A)
2437870 + 0 singletons (6.00% : N/A)

**LASS6**

36133532 + 0 mapped (81.94% : N/A)
32507760 + 0 properly paired (73.86% : N/A)
2715814 + 0 singletons (6.17% : N/A)

**LASS7**

52613039 + 0 mapped (81.95% : N/A)
47327874 + 0 properly paired (73.87% : N/A)
4111115 + 0 singletons (6.42% : N/A)

**LASS8**

32318677 + 0 mapped (79.30% : N/A)
28910540 + 0 properly paired (71.08% : N/A)
2601733 + 0 singletons (6.40% : N/A)

**Table S6: Detailed posterior parameters of the two best models for each species. The best model is indicated in the first row, the second-best in the second row. All other details match Table S2.**

| | Model | k | n | LogL Best | AIC Best | NeCb | NeCe | NeC | NeWe | NeW | mCW | mWC | Ts | Tb | Te | O | theta | mCW2 | mWC2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***EGGPLANT*** | **IM_C_E_E** | **10** | **41** | **-412,9164** | **845,8328** | | **0,086684** | **1,408605** | **0,03403** | **6,775008** | **0,103585** | **0,986315** | **0,752629** | **3,54E-06** | **0,243402** | **0,936605** | **2414,601** | | |
| | IM_E_E | 9 | 39 | -420,5433 | 859,0865 | | 0,235264 | 1,023546 | 0,112459 | 5,721663 | 0,100915 | 1,114656 | 0,574063 | | 0,17156 | 0,939334 | 2628,701 | | |
| ***PEPPER*** | **IM_C_BcCw_E** | **9** | **48** | **-1171,982** | **2361,965** | **0,069361** | | **1,07489** | | **0,319686** | **2,37554** | **1,302736** | **8,524191** | **0,128392** | **0,022028** | **0,999809** | **17203,26** | | |
| | IM_C_E_E | 10 | 45 | -1172,482 | 2364,963 | | 0,023509 | 0,322313 | 4,256316 | 0,540368 | 2,307482 | 1,196217 | 7,90063 | 0,041224 | 0,126791 | 0,999991 | 14914,2 | | |
| ***TOMATO*** | **IM2_BcCw_E** | **10** | **38** | **-2239,996** | **4499,992** | **0,182143** | | **2,498547** | | **1,800286** | **2,59E-47** | **0,051041** | **0,933872** | **0,040991** | | **0,976227** | **3398,733** | **2,203375** | **0,323114** |
| | IM2_C_BcCw_I | 11 | 43 | -2240,001 | 4502,001 | 0,182096 | | 2,498646 | | 1,800134 | 6,16E-06 | 0,050947 | 5,18E-05 | 0,933773 | 0,041016 | 0,976227 | 3398,666 | 2,20212 | 0,323366 |

| | |
|---|---|
| ***k*** | Number of parameters of the model |
| **n** | Number of finished inferences of the 50 independent runs |
| ***LogL*** | Maximum log likelihood value across the 50 independent runs |
| **AIC** | Akaike Information Criterion (AIC), calculated as 2**k* - 2**logL* |
| **Theta** | THETA=4*Nref*U*L -> I want Nref = THETA_ref / (4*U*L) |
| **NeC** | Size of Crop population after exponential growth (relative to Nref). |
| **NeW** | Size of Wild population after exponential growth (relative to Nref). |
| **NeCb** | Size of Crop bottleneck population after split (relative to Nref). |
| **Ts** | Duration in generation of the first epoch (from split to the next epoch)(relative to Nref) |
| **Tb** | Duration of the second epoch after a bottleneck or a change in demographic dynamic (from the end of epoch 1 to the next epoch)(relative to Nref) |
| **Te** | Duration of the third epoch after a change in demographic dynamic (from the end of epoch 2 to the present)(relative to Nref) |
| **mWC** | Migration rate from Crop population to Wild population |
| **mCW** | Migration rate from Wild population to Crop population |

**Ne1=N1/Nref -> I want the effective population size: N1 = Ne1 * Nref*
***T = generations / (2*Nref) -> I want the time of event: gen = T * 2 * Nref*

**Table S7 : Biological conversion of the estimated parameters for the two best models for each species. All ðaði output parameters are given in the white backgrounded table. All parameter conversions or estimates are given in a range of two possible mutation rate (min=5.20x10-09 and max=1x10-08) in the yellow backgrounded table. For each parameter, the best estimate and its standard deviation obtained by bootstrap is provided. The estimated effective size is given as population size and not as ratio and the duration are estimated in generation (in annual plants: 1 generation = 1 year). All other details match Table S2.**

| Mutation Rate per generation | |
|---|---|
| Minimum | 5,20E-09 |
| Maximum | 1,00E-08 |

***EGGPLANT***

Lgenome= 19468437,19

**BEST MODEL**

**IM_C_E_E**

| | parameter | SD | | Max estimates | Max SD | Min estimates | Min SD |
|---|---|---|---|---|---|---|---|
| **theta** | 2414,60137 | 64,98215 | **Nref=** | 3100,66 | 83,45 | 5962,8106 | 160,4722 |
| **NeCe** | 0,08668 | 0,01832 | **NeCe=** | 268,78 | 56,7999448 | 516,88 | 109,23 |
| **NeC** | 1,40860 | 0,17862 | **NeC=** | 378,60 | 48,009345 | 728,08 | 92,33 |
| **NeWe** | 0,0340 | 0,0053 | **NeWe=** | 105,51 | 16,315905 | 202,91 | 31,38 |
| **NeW** | 6,77501 | 0,50433 | **NeW=** | 714,86 | 53,2139037 | 1374,73 | 102,33 |
| **Ts** | 0,75263 | 0,08466 | **genS=** | 2333,65 | 262,49 | 4487,79 | 504,79 |
| **Tb** | 0,00000 | 0,02374 | **genB=** | 0,01 | 73,62 | 0,02 | 141,57 |
| **Te** | 0,2434 | 0,0131 | **genE=** | 754,71 | 40,73 | 1451,36 | 78,32 |
| **mCW** | 0,10358 | 0,01564 | **TIMESPLIT=** | 3088,37 | 336,11 | 5939,17 | 724,69 |
| **mWC** | 0,98631 | 0,03828 | | | | | |
| **mCW2** | | | | | | | |
| **mWC2** | | | | | | | |

**SECOND BEST MODEL**

**IM_E_E**

| | parameter | SD | | Max estimates | Max SD | Min estimates | Min SD |
|---|---|---|---|---|---|---|---|
| **theta** | 2628,70073 | 71,21550 | **Nref=** | 3375,5929 | 91,4499 | 6491,5248 | 175,8653 |
| **NeCe** | 0,23526 | 0,02146 | **NeCe=** | 794,1566 | 72,4446 | 1527,2242 | 139,3165 |
| **NeC** | 1,02355 | 0,11490 | **NeC=** | 812,8562 | 91,2513 | 1563,1849 | 175,4833 |
| **NeWe** | 0,11246 | 0,00671 | **NeWe=** | 379,6166 | 22,6355 | 730,0320 | 43,5297 |
| **NeW** | 5,72166 | 1,17605 | **NeW=** | 2172,0385 | 446,4498 | 4176,9971 | 858,5574 |
| **Ts** | 0,57406 | 0,04732 | **genS=** | 1937,8020 | 159,7393 | 3726,5424 | 307,1909 |
| **Tb** | | | | | | | |
| **Te** | 0,17156 | 0,01603 | **genE=** | 579,1177 | 54,1076 | 1113,6878 | 104,0530 |
| **mCW** | 0,10092 | 0,01567 | **TIMESPLIT=** | 2516,9197 | 159,7393 | 4840,2302 | 411,2440 |
| **mWC** | 1,11466 | 0,05016 | | | | | |
| **mCW2** | | | | | | | |
| **mWC2** | | | | | | | |

***PEPPER***

Lgenome= 18401317,93

**BEST MODEL**

**IM_C_BcCw_E**

| | parameter | SD | | Max estimates | Max SD | Min estimates | Min SD |
|---|---|---|---|---|---|---|---|
| **theta** | 17203,25 | 6288,64601 | **Nref =** | 23372,32 | 8543,74 | 44946,7668 | 16430,2770 |
| **NeCe** | 0,06936 | 0,00817 | **NeCe =** | 1621,13 | 191,002097 | 3117,55 | 367,31 |
| **NeC** | 0,06936 | 0,01194 | **NeC =** | 112,44 | 19,3503786 | 216,24 | 37,21 |
| **NeWe** | | | | | | | |
| **NeW** | 0,31969 | 0,63761 | **NeW =** | 7471,80 | 14902,3228 | 14368,84 | 28658,31 |
| **Ts** | 8,52419 | 3,52031 | **genS =** | 199230,10 | 82277,83 | 383134,81 | 158226,60 |
| **Tb** | 0,1283919 | 0,01753 | **genB =** | 3000,82 | 409,74 | 5770,80 | 787,95 |
| **Te** | 0,02203 | 0,00645 | **genE =** | 514,84 | 150,86 | 990,08 | 290,11 |
| **mCW** | 2,37554 | 1,37081 | **TIME SPLIT =** | 202745,76 | 82687,57 | 389895,69 | 159304,67 |
| **mWC** | 1,30274 | 0,17065 | | | | | |
| **mCW2** | | | | | | | |
| **mWC2** | | | | | | | |

**SECOND BEST MODEL**

**IM_C_E_E**

| | parameter | SD | | Max estimates | Max SD | Min estimates | Min SD |
|---|---|---|---|---|---|---|---|
| **theta** | 14914,20 | 322,34575 | **Nref =** | 20262,41 | 437,94 | 38966,1705 | 842,1892 |
| **NeCe** | 0,02351 | 0,00303 | **NeCe =** | 476,36 | 61,3296095 | 916,07 | 117,94 |
| **NeC** | 0,32231 | 0,20485 | **NeC =** | 153,54 | 97,5814891 | 295,26 | 187,66 |
| **NeWe** | 4,25632 | 36,62907 | **NeWe =** | 86243,22 | 742193,253 | 165852,34 | 1427294,72 |
| **NeW** | 0,54037 | 2,22966 | **NeW =** | 46603,04 | 192293,439 | 89621,22 | 3182388,39 |
| **Ts** | 7,90063 | 2,26767 | **genS =** | 160085,79 | 45948,54 | 307857,28 | 88362,57 |
| **Tb** | 0,04122404 | 0,03338 | **genB =** | 835,30 | 676,37 | 1606,34 | 1300,72 |
| **Te** | 0,12679 | 0,19675 | **genE =** | 2569,09 | 3986,65 | 4940,56 | 7666,64 |
| **mCW** | 2,30748 | 0,77964 | **TIME SPLIT =** | 163490,18 | 46624,91 | 314404,19 | 97329,92 |
| **mWC** | 1,19622 | 0,17979 | | | | | |
| **mCW2** | | | | | | | |
| **mWC2** | | | | | | | |

***TOMATO***

Lgenome= 20160440,23

**BEST MODEL**

| | **IM2_BcCw_E** | | | Max | | Min | |
|---|---|---|---|---|---|---|---|
| | parameter | SD | | estimates | SD | estimates | SD |
| **theta** | 3398,73313 | 702,473286 | **Nref =** | 4214,61 | 871,10 | 8105,0130 | 1675,1992 |
| **NeCe** | 0,18214298 | 0,70602529 | **NeCe =** | 767,66 | 2975,61899 | 1476,27 | 5722,34 |
| **NeC** | 2,49854658 | 15,1003881 | **NeC =** | 1918,04 | 11591,9795 | 3688,53 | 22292,27 |
| **NeWe** | | | | | | | |
| **NeW** | 1,80028572 | 3,57078317 | **NeW =** | 7587,50 | 15049,447 | 14591,34 | 28941,24 |
| **Ts** | 0,93387 | 0,19608 | **genS =** | 3935,90 | 826,41 | 7569,04 | 1589,24 |
| **Tb** | 0,04099 | 0,09558618 | **genB =** | 172,76 | 402,86 | 332,23 | 774,73 |
| **Te** | | | | | | | |
| **mCW** | 2,5926E-47 | 3,64613858 | ***TIME SPLIT =*** | *4108,66* | *1229,26* | 7901,28 | 2363,97 |
| **mWC** | 0,05104093 | 0,37011595 | | | | | |
| **mCW2** | 2,20337528 | 11,1562179 | | | | | |
| **mWC2** | 0,32311394 | 0,354896 | | | | | |

**SECOND BEST MODEL**

| | **IM2_C_BcCw_E** | | | *Max* | | *Min* | |
|---|---|---|---|---|---|---|---|
| | parameter | SD | | estimates | SD | estimates | SD |
| **theta** | 3398,66625 | 245,609108 | **Nref =** | 4214,52 | 304,57 | 8104,8536 | 585,7080 |
| **NeCb** | 0,18209596 | 1,05062378 | **NeCe =** | 767,45 | 4427,87898 | 1475,86 | 8515,15 |
| **NeC** | 2,49864615 | 17,4042873 | **NeC =** | 1917,58 | 13356,8813 | 3687,65 | 25686,31 |
| **NeWe** | | | | | | | |
| **NeW** | 1,80013408 | 3,14756178 | **NeW =** | 7586,71 | 13265,4742 | 14589,82 | 25510,53 |
| **Ts** | 5,1765E-05 | 5,27405816 | **genS =** | 0,22 | 22227,64 | 0,42 | 42745,47 |
| **Tb** | 0,93377 | 4,45849164 | **genB =** | 3935,41 | 18790,42 | 7568,09 | 36135,42 |
| **Te** | 0,04101572 | 0,2096424 | **genE =** | 172,86 | 883,54 | 332,43 | 1699,12 |
| **mCW** | 6,1579E-06 | 4,55851452 | ***TIME SPLIT =*** | *4108,49* | *41018,06* | 7900,94 | 80580,01 |
| **mWC** | 0,05094692 | 0,26269995 | | | | | |
| **mCW2** | 2,20211985 | 18,07361 | | | | | |
| **mWC2** | 0,32336635 | 0,59381701 | | | | | |

# APPENDIX 3: Supplementary tables related to the chapter 3

**Table S1: Detailed data about the studied accessions. The species and the location of origins are listed in separated tables for the three accessions: a. the eggplant accessions, b. the pepper accessions and c. the tomato accessions.**

| S1*a* | Name | Species | Clade | POP | Country | IDs |
|---|---|---|---|---|---|---|
| | MM0014 | *S. melongena* | | Crop | Greece | |
| | MM0620 | *S. melongena* | | Crop | India | |
| | MM0669 | *S. insanum* | | Wild | India | |
| | MM0686 | *S. insanum* | | Wild | Indonésie | |
| | MM0693 | *S. insanum* | | Wild | Sri Lanka | |
| | MM0709 | *S. insanum* | | Wild | Malaysia | |
| | MM0710 | *S. insanum* | | Wild | Thailand | |
| | MM0730 | *S. melongena* | | Crop | India | |
| | MM1192 | *S. insanum* | | Wild | Madagascar | |
| | MM1407 | *S. insanum* | | Wild | Sri Lanka | |
| | MM1572 | *S. melongena* | | Crop | Thailand | |
| | MM1678 | *S. insanum* | | Wild | Thailand | |
| | MM1789 | *S. insanum* | | Wild | Vietnam | |
| | MM1803 | *S. melongena* | | Crop | Egypt | |
| | MM1826 | *S. melongena* | | Crop | China | |
| | MM1831 | *S. melongena* | | Crop | China | |
| | MM1838 | *S. insanum* | | Wild | Vietnam | |
| | MM1900 | *S. insanum* | | Wild | Thailand | |
| **S1*b*** | **Name** | **Species** | **Clade** | **POP** | **Country** | **IDs** |
| | PM0076 | *C. annuum* | | Crop | France | |
| | PM0549 | *C. annuum* | | Crop | Hungary | |
| | PM0568 | *C. annuum* | | Crop | Italy | |
| | PM0609 | *C. annuum* | | Crop | Mexico | |
| | PM0647 | *C. annuum glab.* | | Crop | Mexico | |
| | PM0702 | *C. annuum* | | Crop | Mexico | |
| | PM0828 | *C. annuum glab.* | | Crop | NA | |
| | PM0910 | *C. annuum* | | Crop | Turquey | |
| | PM0952 | *C. frutescens* | | Wild | Guatemala | |
| | PM1093 | *C. chinense* | | Wild | Mexico | |
| | PM1100 | *C. annuum* | | Crop | Cuba | |
| | PM1219 | *C. frutescens* | | Wild | Nepal | |
| | PM1565 | *C. annuum* | | Crop | China | |
| | PM1573 | *C. annuum* | | Crop | Sudan | |
| | PM1621 | *C. chinense* | | Wild | Cameroun | |
| **S1*c*** | **Name** | **Species** | **Clade** | **POP** | **Country** | **IDs** |
| | LACMVSel | *S. peruvianum* | peruvianum | Wild | ? | *LACMVSel* |
| | LA1274 | *S. peruvianum* | peruvianum | Wild | Peru | LA1274 |
| | LA1283 | *S. corneliomulleri* | peruvianum | Wild | Peru | LA1283 |
| | LA1358 | *S. huaylasense* | peruvianum | Wild | Peru | LA1358 |
| | LA1365 | *S. huaylasense* | peruvianum | Wild | Peru | LA1365 |
| | LA1552 | *S. corneliomulleri* | peruvianum | Wild | Peru | LA1552 |
| | LASC1 | *S. lycopersicum* | lycopersicum | Crop | Modern Cultivars | *Levovil* |
| | LASC10 | *S. lycopersicum* | lycopersicum | Crop | Ecuador | LA0409 |
| | LASC2 | *S. lycopersicum* | lycopersicum | Crop | Modern Cultivars | *Stupicke Polni Rane* |
| | LASC3 | *S. lycopersicum* | lycopersicum | Crop | Modern Cultivars | *Plovdiv 24A* |
| | LASC4 | *S. lycopersicum* | lycopersicum | Crop | Ecuador | LA1420 |
| | LASC5 | *S. lycopersicum* | lycopersicum | Crop | Modern Cultivars | *Criollo* |
| | LASC8 | *S. lycopersicum* | lycopersicum | Crop | Modern Cultivars | *Ferum* |
| | LASC9 | *S. lycopersicum* | lycopersicum | Crop | Guatemala | LA0767 |

Table S3: Mapping summary statistics on mapped, properly paired and singletons of all the studied accessions aligned to their reference genome. The three species are detailed in separated tables a. the eggplant accessions, b. the pepper accessions and c. the tomato accessions.

| | **(a) Eggplant** | | **(b) Pepper** | | **(c) Tomato** | |
|---|---|---|---|---|---|---|
| | **MM0014** | | **PM0076** | | **LA1274** | |
| mapped | 31456707 | 80,25% | 34351236 | 73,04% | 21194891 | 81,25% |
| properly paired | 28100798 | 72,02% | 30761560 | 65,57% | 18919984 | 72,68% |
| singletons | 2575196 | 6,60% | 2726499 | 5,81% | 1661057 | 6,38% |
| | **MM0620** | | **PM0549** | | **LA1283** | |
| mapped | 27511008 | 81,05% | 38684906 | 68,97% | 25934953 | 81,13% |
| properly paired | 24958758 | 73,86% | 33682080 | 60,19% | 23134048 | 72,53% |
| singletons | 1984248 | 5,87% | 3894904 | 6,96% | 2050713 | 6,43% |
| | **MM0669** | | **PM0568** | | **LA1358** | |
| mapped | 22070060 | 80,33% | 22848497 | 73,28% | 21563816 | 81,91% |
| properly paired | 19840690 | 72,54% | 20372920 | 65,50% | 19024708 | 72,43% |
| singletons | 1724537 | 6,30% | 1962187 | 6,31% | 2000279 | 7,62% |
| | **MM0686** | | **PM0609** | | **LA1365** | |
| mapped | 32088427 | 79,94% | 25152847 | 72,97% | 25387084 | 80,70% |
| properly paired | 28979926 | 72,52% | 22244024 | 64,70% | 22402578 | 71,35% |
| singletons | 2411486 | 6,03% | 2265063 | 6,59% | 2393309 | 7,62% |
| | **MM0693** | | **PM0647** | | **LA1552** | |
| mapped | 34481776 | 76,66% | 39553334 | 67,87% | 31990784 | 83,36% |
| properly paired | 30847302 | 68,86% | 34391660 | 59,15% | 28784368 | 75,17% |
| singletons | 2779315 | 6,20% | 4002601 | 6,88% | 2486926 | 6,49% |
| | **MM0709** | | **PM0702** | | **LACMVSel** | |
| mapped | 24708645 | 80,92% | 25549332 | 71,72% | 25447071 | 82,29% |
| properly paired | 22230632 | 73,17% | 22715908 | 63,93% | 22721950 | 73,63% |
| singletons | 1807917 | 5,95% | 2197797 | 6,19% | 1977920 | 6,41% |
| | **MM0710** | | **PM0828** | | **LASC10** | |
| mapped | 26753971 | 80,16% | 20214606 | 74,58% | 16150482 | 81,47% |
| properly paired | 23919582 | 72,11% | 18170110 | 67,21% | 14660212 | 74,29% |
| singletons | 1905102 | 5,74% | 1517960 | 5,62% | 1028254 | 5,21% |
| | **MM0730** | | **PM0910** | | **LASC1** | |
| mapped | 22920320 | 81,47% | 28314720 | 71,97% | 23577221 | 84,88% |
| properly paired | 20687624 | 73,86% | 25392088 | 64,70% | 21490774 | 77,72% |
| singletons | 1673126 | 5,97% | 2293676 | 5,84% | 1436160 | 5,19% |
| | **MM1192** | | **PM0952** | | **LASC2** | |
| mapped | 32686082 | 81,31% | 21298147 | 75,37% | 18644911 | 85,10% |
| properly paired | 29054800 | 72,77% | 19083514 | 67,70% | 17040114 | 78,12% |
| singletons | 2345442 | 5,87% | 1723731 | 6,12% | 1077622 | 4,94% |
| | **MM1407** | | **PM1093** | | **LASC3** | |
| mapped | 23825171 | 80,41% | 22679090 | 74,71% | 19515571 | 84,22% |
| properly paired | 21619368 | 73,29% | 20035388 | 66,17% | 17820406 | 77,23% |
| singletons | 1674289 | 5,68% | 1999307 | 6,60% | 1170354 | 5,07% |
| | **MM1572** | | **PM1100** | | **LASC4** | |
| mapped | 22589726 | 74,69% | 27838757 | 72,72% | 13095682 | 84,36% |
| properly paired | 20419022 | 67,80% | 24681030 | 64,63% | 12015946 | 77,74% |
| singletons | 1667793 | 5,54% | 2475860 | 6,48% | 728235 | 4,71% |
| | **MM1678** | | **PM1219** | | **LASC5** | |
| mapped | 22114684 | 78,32% | 21352302 | 74,76% | 17897038 | 84,61% |
| properly paired | 19646196 | 69,97% | 19098706 | 67,04% | 16433892 | 78,04% |
| singletons | 1781309 | 6,34% | 1775325 | 6,23% | 951286 | 4,52% |
| | **MM1789** | | **PM1565** | | **LASC8** | |
| mapped | 24731443 | 78,93% | 25595710 | 73,55% | 7419023 | 76,23% |
| properly paired | 22036914 | 70,74% | 22736268 | 65,49% | 6664180 | 68,76% |
| singletons | 1909981 | 6,13% | 2227599 | 6,42% | 496526 | 5,12% |

| | (a) Eggplant | | (b) Pepper | | (c) Tomato | |
|---|---|---|---|---|---|---|
| | **MM1803** | | **PM1573** | | **LASC9** | |
| mapped | 25541703 | 81,54% | 25258354 | 75,19% | 13514836 | 82,72% |
| properly paired | 23083828 | 74,03% | 22615560 | 67,48% | 12227882 | 75,15% |
| singletons | 1868910 | 5,99% | 2054664 | 6,13% | 926345 | 5,69% |
| | **MM1826** | | **PM1621** | | | |
| mapped | 29975267 | 81,04% | 28705000 | 73,98% | | |
| properly paired | 27060626 | 73,50% | 25563916 | 66,05% | | |
| singletons | 2232206 | 6,06% | 2402262 | 6,21% | | |
| | **MM1831** | | | | | |
| mapped | 33789446 | 81,55% | | | | |
| properly paired | 30613360 | 74,22% | | | | |
| singletons | 2417573 | 5,86% | | | | |
| | **MM1838** | | | | | |
| mapped | 32817699 | 79,13% | | | | |
| properly paired | 29530432 | 71,53% | | | | |
| singletons | 2550808 | 6,18% | | | | |
| | **MM1900** | | | | | |
| mapped | 23828994 | 80,58% | | | | |
| properly paired | 21317108 | 72,43% | | | | |
| singletons | 1935638 | 6,58% | | | | |

Table S3

| **(a) Eggplant** | | **(b) Pepper** | | **(c) Tomato** | |
|---|---|---|---|---|---|
| Accession | mapped | Accession | mapped | Accession | mapped |
| **MM0014** | 80,25% | **PM0076** | 73,04% | **LA1274** | 81,25% |
| **MM0620** | 81,05% | **PM0549** | 68,97% | **LA1283** | 81,13% |
| **MM0669** | 80,33% | **PM0568** | 73,28% | **LA1358** | 81,91% |
| **MM0686** | 79,94% | **PM0609** | 72,97% | **LA1365** | 80,70% |
| **MM0693** | 76,66% | **PM0647** | 67,87% | **LA1552** | 83,36% |
| **MM0709** | 80,92% | **PM0702** | 71,72% | **LACMVSel** | 82,29% |
| **MM0710** | 80,16% | **PM0828** | 74,58% | **LASC10** | 81,47% |
| **MM0730** | 81,47% | **PM0910** | 71,97% | **LASC1** | 84,88% |
| **MM1192** | 81,31% | **PM0952** | 75,37% | **LASC2** | 85,10% |
| **MM1407** | 80,41% | **PM1093** | 74,71% | **LASC3** | 84,22% |
| **MM1572** | 74,69% | **PM1100** | 72,72% | **LASC4** | 84,36% |
| **MM1678** | 78,32% | **PM1219** | 74,76% | **LASC5** | 84,61% |
| **MM1789** | 78,93% | **PM1565** | 73,55% | **LASC8** | 76,23% |
| **MM1803** | 81,54% | **PM1573** | 75,19% | **LASC9** | 82,72% |
| **MM1826** | 81,04% | **PM1621** | 73,98% | | |
| **MM1831** | 81,55% | | | | |
| **MM1838** | 79,13% | | | | |
| **MM1900** | 80,58% | | | | |

Table S4: Summary of numeric results of expressed genes and SNPs detected with the variant calling for the three species. All details are given before and after filtering for paralogs.

| | | *No filter* | *Filtered* | *Minimum Filter*** | *Percentage quality* | *Percentage Filtered* |
|---|---|---|---|---|---|---|
| **Eggplant** | | *727629* | 112773 | 416927 | 57,30% | 27,05% |
| **Pepper** | *SNPs* | *1061975* | 213683 | 597667 | 56,28% | 35,75% |
| **Tomato** | | *2912381* | 950036 | 1945141 | 66,79% | 48,84% |

| | | *Raw mapped* | *Filtered* | *CDS* | *Percentage RC* | *Percentage Filtered* |
|---|---|---|---|---|---|---|
| **Eggplant** | | *33209* | 17545 | 34396 | 96,55% | 51,01% |
| **Pepper** | *genes* | *34610* | 18047 | 35336 | 97,95% | 51,07% |
| **Tomato** | | *34297* | 19628 | 35768 | 95,89% | 54,88% |

** Paralog filter*

**(min-meanDP 10, minQ 20, remove-filtered-geno-all, remove-filtered-all, remove-indels)

Table S6: Detailed per chromosome and global mean of nucleotide diversity for both populations of our three species.

| | Chromosomes: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Global mean | t.test pvalues | sd | ks.test (distribution greater) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Eggplant** | **Pi Crop** | 3,29E-04 | 5,45E-04 | 2,67E-04 | 2,28E-04 | 6,82E-04 | 4,12E-04 | 4,34E-04 | 6,14E-04 | 6,34E-04 | 4,09E-04 | 2,84E-04 | 3,78E-04 | **4,14E-04** | 9,66E-06 | 3,80E-03 | <2,2e-16 |
| | **Pi Wild** | 6,75E-04 | 6,14E-04 | 4,07E-04 | 3,98E-04 | 7,00E-04 | 7,16E-04 | 5,72E-04 | 9,67E-04 | 9,72E-04 | 7,50E-04 | 6,14E-04 | 4,30E-04 | **6,52E-04** | | 5,14E-03 | |
| **Pepper** | **Chromosomes:** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | | | | |
| | **Pi Crop** | 5,22E-04 | 6,57E-04 | 6,66E-04 | 5,47E-04 | 7,25E-04 | 5,81E-04 | 5,67E-04 | 6,71E-04 | 6,87E-04 | 6,57E-04 | 7,08E-04 | 5,98E-04 | **6,40E-04** | <2,2e-16 | 3,04E-03 | <2,2e-16 |
| | **Pi Wild** | 2,42E-03 | 2,23E-03 | 2,84E-03 | 2,41E-03 | 2,27E-03 | 3,16E-03 | 2,23E-03 | 2,22E-03 | 2,57E-03 | 2,60E-03 | 3,20E-03 | 2,56E-03 | **2,61E-03** | | 1,13E-02 | |
| **Tomato** | **Chromosomes:** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | | | | |
| | **Pi Crop** | 1,34E-04 | 2,19E-04 | 1,72E-04 | 2,52E-04 | 2,98E-04 | 1,47E-04 | 1,90E-04 | 2,24E-04 | 4,05E-04 | 1,41E-04 | 3,19E-04 | 2,65E-04 | **2,20E-04** | <2,2e-16 | 8,42E-04 | <2,2e-16 |
| | **Pi Wild** | 2,57E-03 | 2,99E-03 | 2,76E-03 | 2,66E-03 | 2,71E-03 | 2,96E-03 | 2,66E-03 | 2,99E-03 | 2,58E-03 | 2,54E-03 | 2,73E-03 | 3,00E-03 | **2,76E-03** | | 4,72E-03 | |

| | PiCrop/ PiWild |
|---|---|
| **Eggplant** | 6,36E-01 |
| **Pepper** | 0,245213392 |
| **Tomato** | 0,0798481 |

**Table S7 :** Summary results from the DESEQ analyses that detected up- and down-regulated levels of gene expression in crop population compare to the wild population. The summary is detailed for each of the three species.

| Species | *Up regulated in Crop* | *Down regulated in Crop* | vs | Total DEG number | percentage DEG | Number Up / total | Number Down / total | CDS Filtered |
|---|---|---|---|---|---|---|---|---|
| **Eggplant** | 3924 | 4420 | Wild | 8344 | 47,6% | 47,0% | 53,0% | 17545 |
| **Pepper** | 381 | 427 | Wild | 808 | 4,5% | 47,2% | 52,8% | 18047 |
| **Tomato** | 2170 | 2778 | Wild | 4948 | 25,2% | 43,9% | 56,1% | 19628 |

**Table S8:** Gene ontology analyses results for the **Pi-shifted genes of the three species. The group A represent the genes more diverse in the crop population and the group B the genes more diverse in the wild population.**

| | EGGPLANT | | | |
|---|---|---|---|---|
| **GROUP A** | | | | |
| **GO category** | **over represented p-value** | **Num. In Group** | **Num. in gene space** | **Molecular Function** |
| 16311 | $1.48\times10^{-02}$ | 2 | 14 | Dephosphorylation |
| 6813 | $2.68\times10^{-02}$ | 1 | 2 | potassium ion transport |
| 9435 | $3.99\times10^{-02}$ | 1 | 3 | NAD biosynthetic process |
| 9607 | $6.56\times10^{-02}$ | 1 | 5 | response to biotic stimulus |
| 9073 | $6.58\times10^{-02}$ | 1 | 5 | aromatic amino acid family biosynthetic process |
| 7018 | $7.26\times10^{-02}$ | 2 | 33 | microtubule-based movement |
| 45454 | $7.32\times10^{-02}$ | 3 | 72 | cell redox homeostasis |
| 6629 | $9.59\times10^{-02}$ | 3 | 81 | lipid metabolic process |
| **GROUP B** | | | | |
| 30150 | $2.62\times10^{-02}$ | 1 | 1 | protein import into mitochondrial matrix |
| 45132 | $2.64\times10^{-02}$ | 1 | 1 | meiotic chromosome segregation |
| 7131 | $2.66\times10^{-02}$ | 1 | 1 | reciprocal meiotic recombination |
| 16125 | $2.66\times10^{-02}$ | 1 | 1 | sterol metabolic process |
| 469 | $2.69\times10^{-02}$ | 1 | 1 | cleavage involved in rRNA processing |
| 9236 | $5.17\times10^{-02}$ | 1 | 2 | cobalamin biosynthetic process |
| 9298 | $5.17\times10^{-02}$ | 1 | 2 | GDP-mannose biosynthetic process |
| 22900 | $5.24\times10^{-02}$ | 1 | 2 | electron transport chain |
| 9245 | $7.65\times10^{-02}$ | 1 | 3 | lipid A biosynthetic process |
| 6221 | $7.76\times10^{-02}$ | 1 | 3 | pyrimidine nucleotide biosynthetic process |
| 9611 | $7.84\times10^{-02}$ | 1 | 3 | response to wounding |
| 6260 | $9.62\times10^{-02}$ | 2 | 20 | DNA replication |

| | PEPPER | | | |
|---|---|---|---|---|
| **GROUP A** | | | | |
| **GO category** | **over represented p-value** | **Num. In Group** | **Num. in gene space** | **Molecular Function** |
| 6401 | $1.58\times10^{-02}$ | 1 | 1 | RNA catabolic process |
| 16125 | $1.59\times10^{-02}$ | 1 | 1 | sterol metabolic process |
| 19348 | $1.60\times10^{-02}$ | 1 | 1 | dolichol metabolic process |
| 34220 | $1.62\times10^{-02}$ | 1 | 1 | ion transmembrane transport |
| 15986 | $1.69\times10^{-02}$ | 2 | 13 | ATP synthesis coupled proton transport |
| 6207 | $2.92\times10^{-02}$ | 1 | 2 | 'de novo' pyrimidine nucleobase biosynthetic process |
| 17004 | $3.02\times10^{-02}$ | 1 | 2 | cytochrome complex assembly |
| 9972 | $3.17\times10^{-02}$ | 1 | 2 | cytidine deamination |
| 42742 | $4.57\times10^{-02}$ | 1 | 3 | defense response to bacterium |
| 50832 | $4.57\times10^{-02}$ | 1 | 3 | defense response to fungus |
| 9073 | $9.19\times10^{-02}$ | 1 | 6 | aromatic amino acid family biosynthetic process |

| | PEPPER | | | |
|---|---|---|---|---|
| **GROUP B** | | | | |
| **9733** | **$9.53\times10^{-07}$** | **7** | **18** | **response to auxin** |
| **7017** | **$6.64\times10^{-04}$** | **3** | **6** | **microtubule-based process** |
| 42753 | $1.02\times10^{-02}$ | 2 | 5 | positive regulation of circadian rhythm |
| 16485 | $1.49\times10^{-02}$ | 2 | 6 | protein processing |
| 6414 | $2.05\times10^{-02}$ | 2 | 7 | translational elongation |
| 34508 | $3.30\times10^{-02}$ | 1 | 1 | centromere complex assembly |
| 2000123 | $3.30\times10^{-02}$ | 1 | 1 | positive regulation of stomatal complex development |
| 6412 | $5.31\times10^{-02}$ | 8 | 124 | translation |
| 6452 | $6.49\times10^{-02}$ | 1 | 2 | translational frameshifting |
| 45901 | $6.49\times10^{-02}$ | 1 | 2 | positive regulation of translational elongation |
| 45905 | $6.49\times10^{-02}$ | 1 | 2 | positive regulation of translational termination |
| 6850 | $6.49\times10^{-02}$ | 1 | 2 | mitochondrial pyruvate transport |
| 7131 | $6.49\times10^{-02}$ | 1 | 2 | reciprocal meiotic recombination |
| 10167 | $6.49\times10^{-02}$ | 1 | 2 | response to nitrate |
| 15706 | $6.49\times10^{-02}$ | 1 | 2 | nitrate transport |
| 43043 | $6.49\times10^{-02}$ | 1 | 2 | peptide biosynthetic process |
| 43085 | $6.49\times10^{-02}$ | 1 | 2 | positive regulation of catalytic activity |
| 42545 | $8.14\times10^{-02}$ | 3 | 31 | cell wall modification |
| 6397 | $8.61\times10^{-02}$ | 2 | 15 | mRNA processing |
| 6741 | $9.58\times10^{-02}$ | 1 | 3 | NADP biosynthetic process |
| 48278 | $9.58\times10^{-02}$ | 1 | 3 | vesicle docking |

| | TOMATO | | | |
|---|---|---|---|---|
| **GROUP A** | | | | |
| **GO category** | **over** | **Num. In** | **Num. in** | **Molecular Function** |
| **9611** | **$4.55\times10^{-03}$** | **2** | **13** | **response to wounding** |
| **19684** | **$8.87\times10^{-03}$** | **1** | **3** | **Photosynthesis, light reaction** |
| 9767 | $1.10\times10^{-02}$ | 1 | 2 | photosynthetic electron transport chain |
| 6351 | $1.62\times10^{-02}$ | 2 | 51 | Transcription, DNA-templated |
| **GROUP B** | | | | |
| 6368 | $1.44\times10^{-02}$ | 1 | 2 | transcription elongation from RNA polymerase II promoter |
| 16570 | $1.44\times10^{-02}$ | 1 | 2 | histone modification |
| 8033 | $5.89\times10^{-02}$ | 2 | 17 | tRNA processing |
| 42742 | $6.819\times10^{-02}$ | 1 | 2 | defense response to bacterium |
| 50832 | $6.819\times10^{-02}$ | 1 | 2 | defense response to fungus |
| 32012 | $7.41\times10^{-02}$ | 1 | 6 | regulation of ARF protein signal transduction |
| 34227 | $7.88\times10^{-02}$ | 1 | 2 | tRNA thio-modification |
| 6450 | $9.15\times10^{-02}$ | 1 | 1 | regulation of translational fidelity |

**Table S9:** Gene ontology analyses results for DEG down- and up- regulated separately for the three

| GO category | over represented p-value | Num. In DEG | Num. in gene space | Molecular Function |
|---|---|---|---|---|
| **EGGPLANT** | | | | |
| **GO Down-regulated in *S.melongena* vs** | | | | |
| 5975 | $2.16\times10^{-09}$ | 89 | 224 | carbohydrate metabolic process |
| 6486 | $8.46\times10^{-05}$ | 25 | 54 | protein glycosylation |
| 6457 | $4.95\times10^{-04}$ | 14 | 26 | protein folding |
| 6886 | $7.93\times10^{-04}$ | 28 | 71 | intracellular protein transport |
| 6260 | $1.67\times10^{-03}$ | 12 | 23 | DNA replication |
| 6270 | $2.16\times10^{-03}$ | 6 | 8 | DNA replication initiation |
| 51225 | $2.52\times10^{-03}$ | 4 | 4 | spindle assembly |
| 16192 | $6.52\times10^{-03}$ | 21 | 56 | vesicle-mediated transport |
| 7018 | $8.62\times10^{-03}$ | 17 | 43 | microtubule-based movement |
| 226 | $1.10\times10^{-02}$ | 4 | 5 | microtubule cytoskeleton organization |
| 7020 | $1.10\times10^{-02}$ | 4 | 5 | microtubule nucleation |
| 6412 | $1.10\times10^{-02}$ | 56 | 188 | translation |
| 32012 | $2.47\times10^{-02}$ | 4 | 6 | regulation of ARF protein signal transduction |
| 6096 | $2.79\times10^{-02}$ | 13 | 34 | glycolytic process |
| 6887 | $2.96\times10^{-02}$ | 9 | 21 | exocytosis |
| 8652 | $3.76\times10^{-02}$ | 3 | 4 | cellular amino acid biosynthetic process |
| 15689 | $3.85\times10^{-02}$ | 3 | 4 | molybdate ion transport |
| 6396 | $4.12\times10^{-02}$ | 13 | 36 | RNA processing |
| 7030 | $4.61\times10^{-02}$ | 2 | 2 | Golgi organization |
| 9082 | $4.76\times10^{-02}$ | 2 | 2 | branched-chain amino acid biosynthetic process |
| 7064 | $4.82\times10^{-02}$ | 2 | 2 | mitotic sister chromatid cohesion |
| 71704 | $4.90\times10^{-02}$ | 2 | 2 | organic substance metabolic process |
| 6303 | $4.96\times10^{-02}$ | 2 | 2 | double-strand break repair via non-homologous end |
| 7059 | $4.98\times10^{-02}$ | 2 | 2 | chromosome segregation |
| 30150 | $4.99\times10^{-02}$ | 2 | 2 | protein import into mitochondrial matrix |
| **GO Up-regulated in *S. melongena* vs *S.*** | | | | |
| 6468 | $2.18\times10^{-14}$ | 273 | 728 | protein phosphorylation |
| 9733 | $2.53\times10^{-06}$ | 28 | 50 | response to auxin |
| 55114 | $3.32\times10^{-06}$ | 241 | 745 | oxidation-reduction process |
| 6950 | $7.20\times10^{-06}$ | 32 | 62 | response to stress |
| 16567 | $9.22\times10^{-06}$ | 24 | 42 | protein ubiquitination |
| 55085 | $2.75\times10^{-05}$ | 110 | 309 | transmembrane transport |
| 48544 | $2.44\times10^{-04}$ | 17 | 30 | recognition of pollen |
| 6355 | $2.57\times10^{-04}$ | 145 | 446 | regulation of transcription, DNA-templated |
| 6952 | $2.65\times10^{-03}$ | 12 | 22 | defense response |
| 42545 | $4.03\times10^{-03}$ | 14 | 28 | cell wall modification |
| 9415 | $4.56\times10^{-03}$ | 5 | 6 | response to water |
| 6810 | $7.22\times10^{-03}$ | 54 | 158 | transport |
| 8152 | $1.05\times10^{-02}$ | 113 | 370 | metabolic process |
| 272 | $1.27\times10^{-02}$ | 5 | 7 | polysaccharide catabolic process |
| 6511 | $1.49\times10^{-02}$ | 17 | 41 | ubiquitin-dependent protein catabolic process |
| 15696 | $1.64\times10^{-02}$ | 3 | 3 | ammonium transport |
| 6812 | $1.66\times10^{-02}$ | 14 | 32 | cation transport |
| 6820 | $1.72\times10^{-02}$ | 4 | 5 | anion transport |
| 8610 | $1.90\times10^{-02}$ | 8 | 15 | lipid biosynthetic process |
| 6367 | $3.51\times10^{-02}$ | 4 | 6 | transcription initiation from RNA polymerase II promoter |
| 8272 | $3.76\times10^{-02}$ | 6 | 11 | sulfate transport |
| 9793 | $3.96\times10^{-02}$ | 4 | 6 | embryo development ending in seed dormancy |
| 6814 | $4.90\times10^{-02}$ | 3 | 4 | sodium ion transport |

**PEPPER**

**GO Down-regulated in *C. annuum* vs Wild (*C. frutescens* & *C. chinense*)**

| GO category | over represented p-value | Num. In DEG | Num. in gene space | Molecular Function |
|---|---|---|---|---|
| 6855 | $4.30\times10^{-03}$ | 5 | 35 | drug transmembrane transport |
| 55114 | $4.44\times10^{-03}$ | 37 | 747 | oxidation-reduction process |
| 7034 | $8.75\times10^{-03}$ | 3 | 14 | vacuolar transport |
| 6508 | $9.29\times10^{-03}$ | 12 | 176 | proteolysis |
| 15986 | $2.00\times10^{-02}$ | 3 | 19 | ATP synthesis coupled proton transport |
| 45087 | $3.02\times10^{-02}$ | 1 | 1 | innate immune response |
| 9236 | $3.32\times10^{-02}$ | 1 | 1 | cobalamin biosynthetic process |
| 9446 | $3.38\times10^{-02}$ | 1 | 1 | putrescine biosynthetic process |

**GO Up-regulated in *C. annuum* vs Wild (*C. frutescens* & *C. chinense*)**

| | | | | |
|---|---|---|---|---|
| 16226 | $2.24\times10^{-02}$ | 2 | 10 | iron-sulfur cluster assembly |
| 5991 | $2.34\times10^{-02}$ | 1 | 1 | trehalose metabolic process |
| 6188 | $2.42\times10^{-02}$ | 1 | 1 | IMP biosynthetic process |
| 6432 | $2.46\times10^{-02}$ | 1 | 1 | phenylalanyl-tRNA aminoacylation |
| 9058 | $2.70\times10^{-02}$ | 5 | 71 | biosynthetic process |
| 8152 | $3.58\times10^{-02}$ | 15 | 375 | metabolic process |
| 8033 | $3.67\times10^{-02}$ | 2 | 13 | tRNA processing |
| 19288 | $4.71\times10^{-02}$ | 1 | 2 | isopentenyl diphosphate biosynthetic process |
| 50992 | $4.71\times10^{-02}$ | 1 | 2 | dimethylallyl diphosphate biosynthetic process |
| 9086 | $4.78\times10^{-02}$ | 1 | 2 | methionine biosynthetic process |
| 6571 | $4.79\times10^{-02}$ | 1 | 2 | tyrosine biosynthetic process |
| 9072 | $4.80\times10^{-02}$ | 1 | 2 | aromatic amino acid family metabolic process |

**TOMATO**

**GO Down-regulated in CROP vs Wild**

| GO category | over represented p-value | Num. In DEG | Num. in gene space | Molecular Function |
|---|---|---|---|---|
| 6412 | $1.32\times10^{-12}$ | 68 | 250 | translation |
| 7018 | $8.43\times10^{-12}$ | 24 | 46 | microtubule-based movement |
| 9725 | $1.47\times10^{-04}$ | 9 | 20 | response to hormone |
| 6075 | $2.75\times10^{-03}$ | 5 | 10 | (1->3)-beta-D-glucan biosynthetic process |
| 5975 | $3.44\times10^{-03}$ | 42 | 245 | carbohydrate metabolic process |
| 6270 | $4.19\times10^{-03}$ | 4 | 7 | DNA replication initiation |
| 34968 | $7.63\times10^{-03}$ | 4 | 8 | histone lysine methylation |
| 7010 | $7.63\times10^{-03}$ | 4 | 8 | cytoskeleton organization |
| 6364 | $8.26\times10^{-03}$ | 7 | 22 | rRNA processing |
| 6468 | $9.37\times10^{-03}$ | 104 | 741 | protein phosphorylation |
| 5985 | $1.91\times10^{-02}$ | 4 | 10 | sucrose metabolic process |
| 9052 | $3.50\times10^{-02}$ | 2 | 3 | pentose-phosphate shunt, non-oxidative branch |

**GO UP-regulated in CROP vs Wild**

| | | | | |
|---|---|---|---|---|
| 55114 | $1.78\times10^{-07}$ | 169 | 821 | oxidation-reduction process |
| 6629 | $5.71\times10^{-06}$ | 32 | 100 | lipid metabolic process |
| 7034 | $9.26\times10^{-04}$ | 7 | 13 | vacuolar transport |
| 8610 | $9.27\times10^{-04}$ | 9 | 20 | lipid biosynthetic process |
| 9733 | $1.44\times10^{-03}$ | 16 | 51 | response to auxin |
| 6952 | $2.14\times10^{-03}$ | 15 | 48 | defense response |
| 6950 | $2.57\times10^{-03}$ | 18 | 63 | response to stress |
| 16311 | $9.41\times10^{-03}$ | 7 | 18 | dephosphorylation |
| 10167 | $2.01\times10^{-02}$ | 2 | 2 | response to nitrate |
| 15706 | $2.01\times10^{-02}$ | 2 | 2 | nitrate transport |
| 6631 | $2.02\times10^{-02}$ | 5 | 12 | fatty acid metabolic process |
| 6388 | $2.16\times10^{-02}$ | 2 | 2 | tRNA splicing, via endonucleolytic cleavage and ligation |
| 9607 | $2.18\times10^{-02}$ | 10 | 35 | response to biotic stimulus |
| 6812 | $2.78\times10^{-02}$ | 10 | 36 | cation transport |
| 42545 | $3.78\times10^{-02}$ | 10 | 38 | cell wall modification |
| 6979 | $4.64\times10^{-02}$ | 14 | 61 | response to oxidative stress |

Table S10: Detailed results from the Fisher test on distribution of the (a) DEG and (b) shifted genes A and B across the different chromosomes of the three species.

(a)

**Eggplant**

| *Chromosome* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *TOTAL* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *DEG* | 956 | 364 | 842 | 623 | 528 | 747 | 574 | 533 | 436 | 655 | 427 | 466 | 7151 |
| *All genes* | 3839 | 1569 | 3092 | 2373 | 1861 | 2878 | 2634 | 2223 | 1642 | 2591 | 1823 | 1911 | 28436 |
| *Fisher test* | 8,18E-01 | 1,80E-01 | **5,43E-02** | 3,55E-01 | **1,92E-02** | 4,60E-01 | **2,69E-03** | 3,48E-01 | 3,25E-01 | 9,09E-01 | 2,02E-01 | 5,79E-01 | |
| | *24,9%* | *23,2%* | *27,2%* | *26,3%* | *28,4%* | *26,0%* | *21,8%* | *24,0%* | *26,6%* | *25,3%* | *23,4%* | *24,4%* | *25,1%* |
| *DEG DOWN* | 418 | 172 | 414 | 312 | 220 | 355 | 286 | 261 | 196 | 304 | 193 | 238 | 3369 |
| *All genes* | 3839 | 1569 | 3092 | 2373 | 1861 | 2878 | 2634 | 2223 | 1642 | 2591 | 1823 | 1911 | 28436 |
| *Fisher test* | 1,29E-01 | 3,57E-01 | **2,88E-02** | 1,04E-01 | 1,00E+00 | 4,91E-01 | 1,86E-01 | 9,19E-01 | 9,07E-01 | 9,00E-01 | 1,55E-01 | 4,92E-01 | |
| | *10,9%* | *11,0%* | *13,4%* | *13,1%* | *11,8%* | *12,3%* | *10,9%* | *11,7%* | *11,9%* | *11,7%* | *10,6%* | *12,5%* | *11,8%* |
| *DEG UP* | 538 | 192 | 428 | 311 | 308 | 392 | 288 | 272 | 240 | 351 | 234 | 228 | 3782 |
| *All genes* | 3839 | 1569 | 3092 | 2373 | 1861 | 2878 | 2634 | 2223 | 1642 | 2591 | 1823 | 1911 | 28436 |
| *Fisher test* | 2,94E-01 | 3,04E-01 | 4,57E-01 | 8,51E-01 | **7,80E-04** | 6,69E-01 | **2,07E-03** | 2,19E-01 | 1,86E-01 | 7,65E-01 | 6,46E-01 | 1,35E-01 | |
| | *14,0%* | *12,2%* | *13,8%* | *13,1%* | *16,6%* | *13,6%* | *10,9%* | *12,2%* | *14,6%* | *13,5%* | *12,8%* | *11,9%* | *13,3%* |

**Pepper**

| *Chromosome* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *TOTAL* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *DEG* | 105 | 92 | 98 | 75 | 66 | 70 | 66 | 65 | 46 | 46 | 51 | 84 | 864 |
| *All genes* | 3788 | 3232 | 3989 | 2440 | 2096 | 2621 | 2075 | 2487 | 1969 | 2104 | 1977 | 2423 | 31201 |
| *Fisher test* | 1,00E+00 | 7,79E-01 | 2,79E-01 | 3,73E-01 | 3,06E-01 | 8,52E-01 | 2,73E-01 | 7,03E-01 | 2,86E-01 | 1,28E-01 | 6,71E-01 | 5,66E-02 | |
| | *2,8%* | *2,8%* | *2,5%* | *3,1%* | *3,1%* | *2,7%* | *3,2%* | *2,6%* | *2,3%* | *2,2%* | *2,6%* | *3,5%* | *2,8%* |
| *DEG DOWN* | 55 | 51 | 53 | 55 | 35 | 39 | 37 | 31 | 26 | 27 | 28 | 44 | 481 |
| *All genes* | 3788 | 3232 | 3989 | 2440 | 2096 | 2621 | 2075 | 2487 | 1969 | 2104 | 1977 | 2423 | 31201 |
| *Fisher test* | 7,26E-01 | 8,81E-01 | 3,35E-01 | 1,17E-02 | 6,48E-01 | 9,34E-01 | 4,09E-01 | 3,05E-01 | 5,07E-01 | 4,08E-01 | 7,77E-01 | 3,07E-01 | |
| | *1,5%* | *1,6%* | *1,3%* | *2,3%* | *1,7%* | *1,5%* | *1,8%* | *1,2%* | *1,3%* | *1,3%* | *1,4%* | *1,8%* | *1,5%* |
| *DEG UP* | 50 | 41 | 45 | 20 | 31 | 31 | 29 | 34 | 20 | 19 | 23 | 40 | 383 |
| *All genes* | 3788 | 3232 | 3989 | 2440 | 2096 | 2621 | 2075 | 2487 | 1969 | 2104 | 1977 | 2423 | 31201 |
| *Fisher test* | 6,40E-01 | 8,02E-01 | 6,45E-01 | 8,12E-02 | 3,09E-01 | 9,26E-01 | 4,74E-01 | 5,11E-01 | 4,58E-01 | 2,15E-01 | 9,16E-01 | 8,81E-02 | |
| | *1,3%* | *1,3%* | *1,1%* | *0,8%* | *1,5%* | *1,2%* | *1,4%* | *1,4%* | *1,0%* | *0,9%* | *1,2%* | *1,7%* | *1,2%* |

**Tomato**

| *Chromosome* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *TOTAL* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *DEG* | 664 | 527 | 524 | 398 | 346 | 414 | 354 | 368 | 335 | 344 | 313 | 335 | 4922 |
| *All genes* | 4439 | 3541 | 3486 | 2846 | 2530 | 2943 | 2557 | 2528 | 2556 | 2574 | 2440 | 2488 | 34928 |
| *Fisher test* | 1,84E-01 | 2,72E-01 | 1,91E-01 | 9,12E-01 | 6,38E-01 | 1,00E+00 | 7,93E-01 | 5,79E-01 | 2,41E-01 | 3,81E-01 | 1,33E-01 | 4,59E-01 | |
| | *15,0%* | *14,9%* | *15,0%* | *14,0%* | *13,7%* | *14,1%* | *13,8%* | *14,6%* | *13,1%* | *13,4%* | *12,8%* | *13,5%* | *14,1%* |
| *DEG DOWN* | 312 | 244 | 241 | 172 | 142 | 168 | 153 | 162 | 150 | 139 | 144 | 135 | 2162 |
| *All genes* | 4439 | 3541 | 3486 | 2846 | 2530 | 2943 | 2557 | 2528 | 2556 | 2574 | 2440 | 2488 | 34928 |
| *Fisher test* | 4,62E-02 | 1,28E-01 | 1,17E-01 | 8,08E-01 | 2,84E-01 | 3,38E-01 | 7,34E-01 | 6,70E-01 | 5,80E-01 | 1,36E-01 | 6,32E-01 | 1,53E-01 | |
| | *7,0%* | *6,9%* | *6,9%* | *6,0%* | *5,6%* | *5,7%* | *6,0%* | *6,4%* | *5,9%* | *5,4%* | *5,9%* | *5,4%* | *6,2%* |
| *DEG UP* | 352 | 283 | 283 | 226 | 204 | 246 | 201 | 206 | 185 | 205 | 169 | 200 | 2760 |
| *All genes* | 4439 | 3541 | 3486 | 2846 | 2530 | 2943 | 2557 | 2528 | 2556 | 2574 | 2440 | 2488 | 34928 |
| *Fisher test* | 9,53E-01 | 8,45E-01 | 6,70E-01 | 9,43E-01 | 7,90E-01 | 4,16E-01 | 9,70E-01 | 6,76E-01 | 2,86E-01 | 9,10E-01 | 1,10E-01 | 8,18E-01 | |
| | *7,9%* | *8,0%* | *8,1%* | *7,9%* | *8,1%* | *8,4%* | *7,9%* | *8,1%* | *7,2%* | *8,0%* | *6,9%* | *8,0%* | *7,9%* |

Table S10: Detailed results from the Fisher test on distribution of the (a) DEG and (b) shifted genes A and B across the different chromosomes of the three species.

(b)

**Eggplant**

| *Chromosome* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *TOTAL* | *total with chr.00* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *A* | 20 | 10 | 13 | 14 | 12 | 19 | 19 | 3 | 10 | 17 | 9 | 17 | 163 | *195* |
| *All Pi* | 1779 | 648 | 1485 | 1049 | 858 | 1187 | 1015 | 894 | 784 | 1102 | 711 | 803 | 12315 | *percentage* |
| *Fisher test* | 0,5751 | 0,5974 | 0,1767 | 0,8887 | 0,7593 | 0,4287 | 0,1590 | 0,0071 | 1,0000 | 0,4966 | 1,0000 | 0,0831 | | *84%* |
| | 1,1% | 1,5% | 0,9% | 1,3% | 1,4% | 1,6% | 1,9% | 0,3% | 1,3% | 1,5% | 1,3% | 2,1% | 1,3% | |
| *Chromosome* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *TOTAL* | *total with chr.00* |
| *B* | 48 | 14 | 26 | 21 | 21 | 43 | 29 | 18 | 29 | 28 | 22 | 12 | 311 | *374* |
| *All Pi* | 1779 | 648 | 1485 | 1049 | 858 | 1187 | 1015 | 894 | 784 | 1102 | 711 | 803 | 12315 | *percentage* |
| *Fisher test* | 0,6874 | 0,6980 | 0,0745 | 0,3515 | 1,0000 | 0,0366 | 0,5342 | 0,4347 | 0,0639 | 0,9204 | 0,3310 | 0,0763 | | *83%* |
| | 2,7% | 2,2% | 1,8% | 2,0% | 2,4% | 3,6% | 2,9% | 2,0% | 3,7% | 2,5% | 3,1% | 1,5% | 2,5% | |

**Pepper**

| *Chromosome* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *TOTAL* | *total with chr.00* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *A* | 25 | 23 | 27 | 14 | 17 | 18 | 10 | 25 | 9 | 12 | 12 | 22 | 214 | *247* |
| *All Pi* | 1641 | 1628 | 1940 | 1082 | 905 | 1188 | 904 | 1324 | 852 | 929 | 862 | 1021 | 14276 | *percentage* |
| *Fisher test* | 0,9146 | 0,9139 | 0,8412 | 0,6958 | 0,3992 | 0,9014 | 0,4741 | 0,2918 | 0,3784 | 0,7787 | 1,0000 | 0,1144 | | *87%* |
| | 1,5% | 1,4% | 1,4% | 1,3% | 1,9% | 1,5% | 1,1% | 1,9% | 1,1% | 1,3% | 1,4% | 2,2% | 1,5% | |
| *Chromosome* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *TOTAL* | *total with chr.00* |
| *B* | 51 | 31 | 69 | 39 | 22 | 41 | 27 | 41 | 37 | 32 | 28 | 40 | 458 | *520* |
| *All Pi* | 1641 | 1628 | 1940 | 1082 | 905 | 1188 | 904 | 1324 | 852 | 929 | 862 | 1021 | 14276 | *percentage* |
| *Fisher test* | 0,8824 | 0,0037 | 0,4143 | 0,4768 | 0,2381 | 0,6693 | 0,8450 | 0,9349 | 0,0926 | 0,7013 | 0,9208 | 0,2364 | | *88%* |
| | 3,1% | 1,9% | 3,6% | 3,6% | 2,4% | 3,5% | 3,0% | 3,1% | 4,3% | 3,4% | 3,2% | 3,9% | 3,2% | |

**Tomato**

| *Chromosome* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *TOTAL* | *total with chr.00* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *A* | 5 | 7 | 5 | 2 | 4 | 8 | 2 | 6 | 8 | 5 | 7 | 3 | 62 | *63* |
| *All Pi* | 2161 | 1777 | 1771 | 1370 | 1156 | 1485 | 1248 | 1188 | 1180 | 1128 | 1077 | 1096 | 16637 | *percentage* |
| *Fisher test* | 0,4397 | 0,8378 | 0,6806 | 0,2367 | 1,0000 | 0,2799 | 0,3238 | 0,4598 | 0,1401 | 0,6158 | 0,1973 | 0,7980 | | *98%* |
| | 0,2% | 0,4% | 0,3% | 0,1% | 0,3% | 0,5% | 0,2% | 0,5% | 0,7% | 0,4% | 0,6% | 0,3% | 0,4% | |
| *Chromosome* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *TOTAL* | *total with chr.00* |
| *B* | 76 | 76 | 61 | 31 | 40 | 66 | 46 | 37 | 29 | 42 | 47 | 50 | 601 | *607* |
| *All Pi* | 2161 | 1777 | 1771 | 1370 | 1156 | 1485 | 1248 | 1188 | 1180 | 1128 | 1077 | 1096 | 16637 | *percentage* |
| *Fisher test* | 0,9023 | 0,1857 | 0,7883 | 0,0090 | 0,8704 | 0,1153 | 0,8752 | 0,4190 | 0,0487 | 0,8054 | 0,2113 | 0,1365 | | *99%* |
| | 3,5% | 4,3% | 3,4% | 2,3% | 3,5% | 4,4% | 3,7% | 3,1% | 2,5% | 3,7% | 4,4% | 4,6% | 3,6% | |

Table S11: Detailed results from the generalized linear models modeling the regression of Delta Pi and

**(i) = glm(formula = as,formula(formula_pi ~ GauLFC + chr), data = Pi_only_Delta)**

| | eggplant | | | | Pepper | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Std.Error | Pr(>\|t\|) | | Estimate | Std.Error | Pr(>\|t\|) | |
| **(Intercept)** | 4,47E-04 | 2,17E-04 | 0,039 | * | 1,36E-03 | 2,76E-05 | <2e-16 | *** |
| **GauLFC** | -2,16E-04 | 1,02E-04 | 3,43E-02 | * | 2,42E-04 | 2,78E-05 | <2e-16 | *** |
| chr.02 | 4,42E-05 | 4,09E-04 | 0,914 | | 4,58E-05 | 3,91E-05 | 0,2411 | |
| chr.03 | 1,98E-05 | 3,20E-04 | 0,9508 | | -1,48E-05 | 3,71E-05 | 0,6901 | |
| chr.04 | 4,57E-05 | 3,53E-04 | 0,8972 | | 1,04E-05 | 4,32E-05 | 0,8098 | |
| chr.05 | -1,42E-04 | 3,82E-04 | 0,7104 | | 2,47E-05 | 4,55E-05 | 0,5881 | |
| chr.06 | 2,00E-04 | 3,41E-04 | 0,5582 | | 7,76E-05 | 4,22E-05 | 0,0657 | . |
| chr.07 | -2,92E-04 | 3,57E-04 | 0,4137 | | 5,06E-05 | 4,54E-05 | 0,2651 | |
| chr.08 | -2,56E-04 | 3,74E-04 | 0,4942 | | -2,25E-05 | 4,13E-05 | 0,5864 | |
| chr.09 | 2,66E-04 | 3,86E-04 | 0,4897 | | -1,16E-04 | 4,70E-05 | 0,0135 | * |
| chr.10 | 1,03E-05 | 3,54E-04 | 0,9767 | | -1,09E-04 | 4,49E-05 | 0,0155 | * |
| chr.11 | -3,59E-04 | 4,00E-04 | 0,3691 | | -1,09E-04 | 4,65E-05 | 0,0196 | * |
| chr.12 | -4,90E-05 | 3,85E-04 | 0,8986 | | -1,33E-05 | 4,44E-05 | 0,7645 | |

**(ii) = glm(formula = as,formula(GauLFC ~ formula_pi + chr), data = Pi_only_Delta)**

| | **eggplant** | | | | **Pepper** | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Std.Error | Pr(>\|t\|) | | Estimate | Std.Error | Pr(>\|t\|) | |
| **(Intercept)** | 0,0872 | 0,0261 | 0,0008 | *** | 0,0066 | 0,0097 | 0,4945 | |
| **GauLFC** | -3,1422 | 1,4843 | 0,0343 | * | 25,1014 | 2,8863 | <2,00E-16 | *** |
| chr.02 | -0,0179 | 0,0494 | 0,7166 | | -0,0219 | 0,0126 | 0,0814 | , |
| chr.03 | 0,0076 | 0,0386 | 0,8445 | | -0,0179 | 0,0120 | 0,1342 | |
| chr.04 | -0,0919 | 0,0426 | 0,0310 | * | -0,0315 | 0,0139 | 0,0234 | * |
| chr.05 | -0,0529 | 0,0460 | 0,2506 | | -0,0292 | 0,0147 | 0,0468 | * |
| chr.06 | -0,0614 | 0,0411 | 0,1355 | | -0,0255 | 0,0136 | 0,0608 | , |
| chr.07 | -0,0466 | 0,0431 | 0,2793 | | -0,0388 | 0,0146 | 0,0080 | ** |
| chr.08 | 0,0059 | 0,0451 | 0,8958 | | -0,0148 | 0,0133 | 0,2662 | |
| chr.09 | -0,0483 | 0,0465 | 0,2988 | | -0,0288 | 0,0151 | 0,0574 | , |
| chr.10 | -0,0310 | 0,0427 | 0,4677 | | -0,0439 | 0,0145 | 0,0024 | ** |
| chr.11 | -0,0593 | 0,0482 | 0,2193 | | -0,0288 | 0,0150 | 0,0551 | , |
| chr.12 | -0,0775 | 0,0464 | 0,0951 | , | -0,0291 | 0,0143 | 0,0415 | * |

# APPENDIX 4: Reviewers comments and major questions for Genome, Biology and Evolution

The reviews will be addressed during the first months of 2019, in order to proceed to a submission in BMC Genomics.

## Referee: 1

### Comments to the Author

This manuscript by Arnoux et al. explores the relationship between expression differences and genetic diversity differences among wild and domesticated species in three groups of Solanaceae.

The overall goals of the paper are interesting and the dataset is rich. However, there are analyses whose absence is notable (synonymous vs. non-synonymous, and expanded GO term list). Additionally, the results highlighted in the discussion (chromosomal binning and GO terms, in particular) draw quite broad conclusions from more restricted results.

I believe the dataset is rich enough that a strengthened analytical framework and discussion that hews more closely to the specific results could produce a strong study. However, in current form, I cannot recommend this manuscript for publication.

### Major questions

1) Chromosomal binning vs. linkage
One aspect of the results and discussion that was not clear was the biological relevance of binning by chromosome. What was the purpose of chromosomal bins specifically? As opposed to linked sites or windows of a certain physical size (related to recombination). Perhaps a window-based approach or something more about linked sites? There was

2) Synonymous vs. non-synonymous
There was no mention of differentiating non-synonymous and synonymous variants. Why not? This seems highly relevant to measures of genetic diversity from transcriptome data in particular. Selection is invoked throughout based on the result of loss of nucleotide diversity alone, but no analyses of amino acid diversity are discussed.

3) Sampling procedure
The greenhouse conditions were not given with much detail. Light regime, moisture, soil conditions, etc. should be provided for an expression data.

4) Effect of tissue density (exp. fruit)
Are the proportions of the three tissue types by mass of original tissue or final concentration of cDNA?
If by mass, then species with less dense fruits would be under sampled for fruit mRNA, which might introduce a consistent bias in the underrepresentation of fruit mRNAs w.r.t. leaf and flower. Much of the results depend on group-specific patterns of evolution, but if the proportions of cDNA were consistently different due to tissue density differences between the groups, this could be a confounding factor (though perhaps one of limited effect). This is also a concern with the possible differentiation between the domesticated and wild species (especially in tomato) where the domesticated species generally have larger and more metabolite-rich fruits. (Again, if these proportions were normalized at the cDNA level, this comment is moot).

5) Use of SNPs and 1:1:1
On p.13 l.37: Were these 12,655 genes just those that were 1:1:1 orthologs and had SNPs in all three? Text is unclear on this.
This appears again in pg. 15 l.10. and pg.17. Why is it relevant to look at a set of 1:1:1 orthologs where all three groups all have SNPs? Why not report the expression and GO overlap regardless of requiring SNPs in all three?

6) Restriction of GO terms to Biological Process. Why restrict the GO term to just Biological Process? No rationale for this was presented.

7) Discussionary points drawn from GO terms
The conclusions drawn about translational mechanisms on page 21 and page 22 is tenuous and drawn only from very general GO term categories of translational genes. The leap from generic GO terms to the very specific functional adaptations given needs additional support. On pg. 25. l. 10 this is called "surprising," but is not specific as to why. An expansion of GO terms beyond Biological Process might be helpful.

8) The discussion of the relationship between genetic variation and expression is also a bit unclear. Particularly the sentence from pg. 23 l.10-16. This seems to skirt around the main result that there is no correlation between diversity and expression shift and try to highlight that a few outliers (?) change the result? Was this actually tested for significance ("p23. l 14)? The remainder of this paragraph seems equivocal on what the expectation was. Were there expectations of a direct correlation between a gene's genetic diversity and expression diversity? Certainly, the works cited would suggest the opposite, given their regulatory complexity. Especially when the overrepresented categories just previously discussed were as putatively pleiotropic as core mitotic function.

## Referee: 2

### Comments to the Author
Review of Arnoux et al, for GBE 2018

This paper describes an analysis of selection related to domestication in three different Solanaceae crop / wild relative pairs: eggplant, tomato, and pepper. It is a potentially interesting topic but I have substantial concerns with the analysis.

### Major questions:
1) The stated goal of the paper is to look at the convergence of selection on gene expression and polymorphisms. While this is an interesting question, it is a challenging one to address correctly. One major challenge is that due to linkage one expects there to be a convergence even if variation in one of the two processes is neutral. If there is a promoter mutation that leads to a change in gene expression that is favorable in the crop species, then that will result both in a difference in gene expression between wild and crop, AND a low pi in the coding region, even if no coding changes were selected for (basically a selective sweep). The converse is also true: if there is neutral variation in gene expression but a coding polymorphism is under selection, this will drive a change in gene expression between crop and domesticated via sweep. Thus, a convergence between these processes is actually what is expected. The challenge, then, is to determine if there is more convergence than expected. This paper does not address this issue.

2) p.6 l. 27. If I understand the design correctly, different tissue types were pooled by weight, rather than RNA amount. This makes differential expression analysis hard to interpret, because developmental differences between species will influence the RNA composition of the pools and confound the analysis. For example, the authors write that the pools contained 20% fruit. If a domesticated species has larger fruit because of (for example) larger vacuoles, or more extra-cellular material, then the pool will likely have less fruit RNA. Fruit expressed genes will then appear to be expressed lower in the domesticate species, not because there cellular expression levels were lower but because the RNA concentration in fruit is lower.

3) p.7 l. 29. I don’t understand why reads2snp paralogous SNP removal was used in this study. The reads2snp pipeline was developed for instances where there reference genomes are not available and therefore collapsing of paralogous genes during de novo assembly is a major concern. Here, with reference genomes available and 150bp PE reads, the vast majority of the reads should map to the correct gene. How many SNPs were removed at this step?

4) p. 8 l. 20 – 36. Two groups of “shifted” genes were defined to be the focus of interest. These are the genes that have the highest nucleotide diversity (pi) in the wild species and lowest in the crop, or vice versa. This somehow assumes that only those genes that are most diverse in the wild (95th quantile) might have been

subject to selection in the crop (or vice versa). This does not make sense; any gene with multiple alleles in the wild could have been selected for in the crop. A more minor point, but why was a quantile threshold used for the upper limit, but a % of maximum value used for the lower threshold?

5) p. 9 l. 40-46. Were p-values for GO categories corrected for multiple testing? It appears not from looking at the code. They need to be.

6) p. 10-11 “Statistical Analyses”. The description an implementation of the GLM does not make sense. See 4 points following.

6A) The authors correctly note that GLMs can be used for non-normal responses. However: it does not make sense to both “Gaussianize” the variables and then to use a GLM to handle non-normal data. If the Gaussianization was successful then the GLM isn’t needed.

6B) To take advantage of the GLM’s ability to handle non-normal data, it is necessary to supply a link function to the non-Gaussian distribution. Checking the code, this was not done. So the model that is fit is actually just a linear regression. This is fine of the data is normal, but then don’t imply that you are doing something that you aren’t.

6C) The manuscript states that the GLM were fit using the lme4 package in R. This does not match the code in the github repository. The code uses the “glm” function which is part of the base “stats” package in R. The function in lme4 is “glmer”. If lmer4 is to be used then one would want to specify fixed and random effects.

6D) The glm models fit log fold change of gene expression as a function of difference in pi (or vice versa). Because LFC can be negative or positive (indicating an decrease or increase in expression in the comparison) there is a directionality implied in these fits that does not make sense. That is, the model tries to associate high delta-pi with positive LFC (or low delta-pi with positive LFC if the regression coefficient is negative). Consider a situation where delta-pi is high, suggestive of possible selection in the crop. If it is gene expression that has been selected for there is no reason to favor the hypothesis that it was high gene expression in the crop that was selected for over the hypothesis that it was low. So if multiple genes in the genome have been subject to selection for differences in gene expression we expect that in some cases it will be for higher gene expression and in some cases it will be low. These would cancel out in the model as written. Perhaps taking the absolute value of LFC first would be a solution.

7) p.13 l 15 – 30. I don’t understand how the Welch test was performed. This is a non-parametric t-test and so I would expect multiple samples per category. However for each species pair isn’t there only one observation of pi for wild and one for crop?

8) P. 22 l.24. Sounds like you are implying that selection acted both on coding changes and gene expression but you haven’t shown that. Selection on either would expect to reduce pi and potentially drag the other via linkage.

# ACKNOWLEDGEMENTS

Je voudrais remercier le jury de la thèse, car ils ont tous accepté de lire les résultats de 3 années de travail et je leur en suis reconnaissante, en tant que rapporteuses : **Joëlle Ronfort** et **Maud Tenaillon**; et en tant qu'examinateurs : **Concetta Burgarella** et **Jérémy Clotault**.

***« Le secret pour avancer est de commencer. Le secret de la mise en route consiste à diviser vos tâches écrasantes complexes en petites tâches gérables et à commencer par la première. »***

***Mark Twain***

***« Une thèse ne se finit pas, elle s'arrête. »***

***Le scientifique inconnu***

Je tiens à remercier les personnes qui m'ont accompagné dans ce long voyage qu'est la thèse. Sans leur soutien, je n'en serais pas là.

Un grand merci à **Christopher Sauvage** de m'avoir permis de joindre ce projet et de travailler au GAFL. Je le remercie aussi pour son encadrement, pour tout le temps et la patience qu'il a investis en moi. Pour ses conseils avisés que j'ai su écouter, le plus souvent, et pour toutes nos conversations scientifiques qui ont fait grandir ma capacité à analyser des données et en tirer les informations intéressantes. J'ai eu de la chance de l'avoir dans la même barque que moi, il a su pointer du doigt la direction à prendre.

Merci à **Mathilde Causse** qui a su aiguiller ma thèse de loin et m'aider à avancer et à continuer la route même pour les échéances importantes. Et je reste persuadée qu'elle cache une âme rock and roll assagie.

Je tiens à remercier tout mon groupe, **Yolande Carretero** pour avoir pris soin de nos plantes et m'avoir appris à différencier les gourmands, **Rebecca Stevens** pour sa passion pour les livres et les belles discussions que nous avons partagé, **Esther Pelpoir** pour les heures de philosophies de comptoir, **Isidore Ambroise Diouf** pour sa cuisine (les fatayas) et les beaux moments musicaux partagés, **Jiantao Zhao** pour sa bonne humeur constante, **Cécile Garchery**, **Karine Pellegrino**, **Estelle Bineau**, **Frédérique Bitton** pour m'avoir permis de trouver ces virgules cachées qui bloquaient mes codes, **Justine Gricourt** pour tout le soutien qu'elle m'a apporté pendant les moments creux de la thèse, et tout particulièrement **Renaud Duboscq** qui a été à l'origine de toutes les données RNAseq que j'ai utilisées lors de ma thèse et avec qui nous avons fait pousser la Yin et le Yang.

Je remercie tout le personnel du GAFL pour leurs sourires et la bonne humeur partagée dans la vie quotidienne (notamment à la cantine). J'ai aussi reçu un soutien considérable par le bio-informaticien de l'unité **Jean-Paul Bouchet** et par le bio-informaticien.2.0 **Jacques Lagnel**, leur aide était précieuse pour moi, ils m'ont apporté une expertise et m'ont permis de développer mes aptitudes à coder.

Je tiens à remercier l'**EIRA** de m'avoir sélectionné et permis d'obtenir une bourse pour aller à Vienne, ainsi que **Beatriz Vicoso** qui a accepté de m'accueillir dans son groupe à l'IST, spécialement **Christelle Fraïsse** avec qui j'ai eu la chance de travailler. Elle m'a enseigné toutes les méthodes d'inférences démographies, et elle a su m'aider, même à distance, dans la rédaction de l'article. C'est une chercheuse remarquable avec qui j'espère pouvoir collaborer dans le futur.

Cette thèse a été suivie par un comité formidable que je tiens à remercier. **Christophe Lemaire** qui a su se rendre disponible et m'a beaucoup aidé à restructurer ma discussion de l'article sur les inférences. **Stéphane de Mita**, qui a relu mes articles, mais surtout qui a su me convaincre d'une manière de maître d'arrêter d'analyser mes données et de commencer à écrire. Je tiens à remercier tout particulièrement, **Ivan Scotti**, qui a fait partie de mon comité mais m'a aussi apporté beaucoup pour mon avancement scientifique personnel. Je le remercie de m'avoir ouvert les portes du JoBip et surtout de m'avoir offert des discussions scientifiques éclairées. Je remercie aussi **Bouchaib Khadari**. Je ne sais pas de quoi demain est fait, mais si je continue la recherche, j'espère faire de la science à côtés de tels chercheurs.

De plus cette thèse a été cofinancée par la **région PACA** en collaboration avec l'entreprise **Gautier Semences**, je tiens à les remercier et tout particulièrement **Clémence Plissonneau** et **Frédéric Moquet** qui ont suivi mon travail et m'ont donné des conseils avisés lors de mes comités de thèse et à l'occasion de rencontres scientifiques.

Je souhaite remercier le **GNIS** et la **FRB** de m'avoir permis de monter à la capitale lors d'une journée parfaite, pour recevoir un prix portant sur la biodiversité.

***« L'amour d'une famille, le centre autour duquel tout gravite et tout brille. »***

***Victor Hugo***

J'ai une famille formidable que je tiens à remercier pour leur soutien continu : mes 4 parents qui m'ont élevé dans l'amour et la bienveillance, m'ont permis de penser par moi-même et d'affronter les épreuves de la vie, ma maman **Pépée** et mon beau-papa vin **Alain**, avec qui je partage d'inoubliables moments de complicité et qui ont toujours été à mon écoute et m'ont soutenu dans mes choix, mon papa **Dédé** et ma belle-maman **Fabienne**, qui me soutiennent malgré la distance et sont toujours de bonne écoute et avec qui nous passons d'incroyable moments en mer, mon frère **Xavier** et son soutien dans tous les moments même les plus difficiles, j'ai grandi avec lui et il m'a toujours poussé à être une meilleur personne, merci à sa petite famille notamment mon petit chameau et rayon de soleil **Lexie**, ma petite sœur **Ornella**, qui m'apporte beaucoup de bonheur.

Merci à mes autres frères et sœurs (**Graziella**, **Alizée**, **Jimmy** et **Teddy**) et leurs familles, proches ou à distance, vous voir vous épanouir et voir grandir mes neveux et nièces me comble de bonheur.

Je remercie les femmes fortes que sont mes deux mamies **Mariette** et **Mado**, et mes papis **Lou** et **Louis** que je n'oublie pas. Les membres de la **B2oBa** family, et tous mes oncles, tantes, cousins et cousines, parce que ma famille est grande en nombre mais en amour aussi.

***« L'amitié double les joies et réduit de moitié les peines. »***

***Francis Bacon***

Comme les amis, c'est la vie, alors merci à **Morgane R.** qui a été l'initiatrice de cette aventure, petite entremetteuse ; à mon amie **Justine C.** dont je suis très fière et avec qui je grandis et deviens une adulte et à son petit **Robin** et ses sourires pour sa 'naine' ; aux piliers pour leur soutien, **Zébra** pour ses gâteaux, et **Pinky** pour son rhum ; à mes amis de toujours : la **Juju**, la **Lala** et son **Yvon**, le p'tit **Moris** et son perroquet **Audrey**, le **Mac Fly** et sa spatule la **Julie**, le **Thib** et sa **Noémie**, le **pierrot** parce qu'ils sont mes potes pour la vie, ma **Poupette** et ses hommes, **Quentin D.** pour son écoute et sa compréhension, **Babel** pour ses goûts littéraires et à **Patate** aussi pour sa calculatrice en terminal.

Je remercie aussi l'EIRA de m'avoir fait rencontrer les autres individus de l'espèce *S. Auriculus*, on s'est reconnu et j'ai eu beaucoup de chance et de soutien de la part de **Modinette** la pilote de flam'kuche, **Alice** ma coloc de rêve, **Norman** mon hébergeur occasionnel, **Gabi** le receleur de licornes, **Nathan** l'abyssal et **Juan** le couillu, autant dire que votre soutien online était une bouffée d'air dans ma noyade finale…

Pendant cette thèse j'ai eu la chance de partir en formation plus de 2 mois cumulés, tout d'abord en Crête à l'**EMBO** et en Allemagne à l'**EVOP** où j'ai rencontré des groupes d'amis scientifiques qui savent garder contact et se soutenir dans les épreuves phylogénétiques ou bio-informatiques. Merci aux groupes Operation Embo-drinker et BadBromance.

Merci à mes anciens collègues de Vienne : (VBC) **Mariana**, **Karin**, **Leni**, **Hagar**, **Ratna**, **Iulia**, **Aleksandra**, **Thomas**, **Geri**, et the **Golden Girls**, pour m'avoir soutenu dans les moments difficiles et m'avoir donné la force de recommencer une thèse, (IST) the **Vicoso lab** pour leur accueil chaleureux. Et à mes amis de Vienne **Marwa S.**, **Lama F.**, **Hussein Z.** pour les valeurs qu'ils m'ont apporté et les saveurs qu'ils m'ont fait découvrir et **Benoit P.** pour avoir relu mon introduction de thèse et faire des commentaires toujours plus décalés.

Merci à mes amis d'Avignon et à l'Explo de nous avoir accueilli à bras ouverts : **Gilles** , **Flo et Loick** de l'explo, **Sandrine** El Racoon, **Isabelle** La magnifique, les petites femmes de mes piliers **Tiff** et **Mél**, ma **primprenelle** parce qu'elle me fait vivre des aventures vers l'infini et l'au-delà, **Cath** pour nos déclarations enflammées et les 32 histoires du soir, **Névine** pour sa générosité et sa compréhension, **Kévin le T.** pour tous les souvenirs incontournables et les rencontres avec des chiens errants, **Max B.** pour le week end à Gre et les belles discussions, **Momy** et ses mouchoirs, **Caro** pour notre girl-team au baby, sans oublier les doctorants d'Avignon (dans l'ordre de disparition) **Léandro** parce que son âme est sœur à la mienne et qu'il m'a fait découvrir le perchoir pour m'évader, **Lucie et Mariem** pour leur musique et leurs sourires, **Hussein** pour ses gâteaux et les meilleurs Mezzés au monde, **Aimeric** pour le coup du siècle au baby : incontournable, **Kévin B.** pour nos discussions philosophiques sur la place de la Femme dans la société et des Hommes dans le monde, **Anne So** pour son soutien de co-thésarde et pour ses prévisions de date de soutenance, **Pierre** pour sa bonne humeur, **Zeid** pour les mémoires oubliées, **Coffi** pour son amitié et sa présence, **et al.**

Merci à mes amis de l'impro, parce que ces aventures imaginaires m'ont permis de prendre de belles bouffées d'air.

Un grand merci à ma grande amie **Élo**, parce que trouver une amie sapin ça n'a pas de prix.

Et enfin je tiens à remercier tout particulièrement mon collègue de bureau, qui m'a supporté, soutenu mais surtout coaché pendant ces 3 ans. Maintenant devenus amis, je suis heureuse que la vie t'aie mis sur mon chemin l'ami **Rémi**, je ne te laisserai pas gagner au baby pour autant.

***« There comes a day when you realize that turning the page is the best feeling in the world - because you realize that the book is much more than the page on which you were stuck. »***

***Stéphanie Arnoux, 2017***

Further ~~reading~~ listening:

**Debussy** - *Arabesque No. 1*, **B. Engerer** – *Noctures de Chopin*, **H. Shore** – *Misty Mountains*, **V. Morrison** – *Brown Eyed* Girl, **Fleetwood Mac** – *Everywhere/Landslide*, **Queen** - *Somebody to Love*, **Fishbach** - *Un autre que moi*, **Parcels** - *Tieduprightnow* , **B. Carlisle** - *The eye*, **A. Franklin** - *Respect*, **Ella & Louis** - *Cheek to Cheek*, **L. Bridges** - *Coming Home*, **Barbara** - *Vienne*, **Fayrouz** - *Ana La Habibi*, **N. Al Saghira** - *Ana Baashaq El Bahr*, **Kokoroko** – *Abusey Junction//We out here*, etc.

## Résumé substantiel de la thèse en français

La domestication des plantes a débuté il y a quelques milliers d'années quand les hommes se sont sédentarisés. Ils ont sélectionné les plantes sauvages portant des caractères phénotypiques d'intérêt pour la consommation et production humaine. Ce processus évolutif a par conséquent modifié le patrimoine génétique des espèces domestiquées. Cette thèse se penche sur les traces génétiques induites par la domestication chez trois espèces de Solanacées : l'aubergine (*Solanum melongena*), le piment (*Capsicum annuum*) et la tomate (*S. lycopersicum*). En effet, si les caractères phénotypiques des plantes cultivées ont été sélectionnés depuis des milliers d'années, les conséquences moléculaires d'une telle sélection restent peu étudiées à l'échelle du génome. Cette étude est basée sur des données de diversité et d'expression de gènes (RNAseq). En utilisant des méthodes comparatives entre des variétés cultivées et leurs espèces sauvages apparentées, j'ai étudié, à l'échelle intra-spécifique, d'une part les histoires démographiques de chacune des espèces, et d'autre part les changements de diversité nucléotidique et d'expression des gènes dus à la domestication. La comparaison de ces trois événements indépendants de domestication, offre l'opportunité de décrypter les changements génétiques qui convergent chez ces trois espèces lors du processus de sélection humaine.

Suite à une introduction qui pose le cadre de cette étude et présente l'état de l'art, le premier chapitre, s'inscrit dans un ouvrage portant sur la génomique des populations d'espèces modèles. Il propose une synthèse des connaissances accumulées en plus d'un siècle de recherche sur l'espèce modèle qu'est la tomate (*S. lycopersicum*). Ce chapitre permet également de compléter le contexte scientifique dans lequel cette thèse s'inscrit, notamment, en retraçant l'importance que les espèces sauvages apparentées ont eu dans l'amélioration de l'adaptabilité des variétés cultivées actuelles.

L'hypothèse du deuxième chapitre révèle la convergence des changements démographiques entre les trois espèces malgré leurs événements indépendants de domestication. L'étude comparée d'inférences de scénarios démographiques a permis de reconstruire l'histoire démographique de chaque espèce cultivée. Ces inférences ont aussi facilité l'estimation des paramètres tels que les flux migratoires entre les espèces sauvages et cultivées, la force des goulots d'étranglement liés à l'intensité de la sélection humaine et la durée des événements de domestication. Ce chapitre permet de démontrer que les changements démographiques liés à la domestication dépendent de l'état de sympatrie ou d'allopatrie des variétés cultivées avec leurs sauvages apparentées. Les connaissances quant à la datation des événements de domestication de nos trois espèces restent très faibles, et les inférences ont permis d'établir des estimations de durée de domestication relativement précise. Ces nouvelles connaissances apportent une plus-value à cette étude pour nos trois espèces et nous invitent à s'interroger sur les différents compartiments du génome qui ont été sélectionnées et modifiées lors de la domestication.

Le troisième chapitre teste l'hypothèse d'une convergence évolutive des changements moléculaires, notamment transcriptionnels, induits par la domestication et l'amélioration moderne. La comparaison des variétés cultivées à leurs espèces sauvages apparentées permet d'évaluer la convergence des mécanismes de régulation et d'adaptation liés à la domestication. C'est en testant la corrélation entre les traces génétiques (diversité nucléotidique) de sélection et les changements d'expression des gènes observés chez les variétés cultivées que l'hypothèse de départ a été validée. Cette analyse montre que la domestication, au-delà même de changements nucléotidiques, a modifié l'expression des gènes chez les trois espèces. L'analyse des gènes orthologues des espèces a confirmé que la domestication a sélectionné des gènes liés aux phénotypes de développement des fruits et la croissance de la plante

alors qu'elle avait, au contraire, contre-sélectionné des gènes liés à la défense des plantes et à leur capacité à tolérer des stress environnementaux.

Enfin, en discussion, je réalise un bilan sur mon projet qui apporte de nombreuses preuves de convergence dues à la domestication et des connaissances utiles pour l'étude de l'histoire des Solanacées. De surcroit, des perspectives d'analyses complémentaires sur la liste de nombreux gènes candidats affectés par la domestication, offrent un potentiel de transversalité, pour l'amélioration des variétés cultivées et pour l'étude plus approfondie des conséquences biologiques et évolutives de la domestication.

## CHAPITRE 1 :

La tomate est une espèce modèle reconnue pour la recherche en génétique et en génomique, sur le développement des fruits et la résistance aux maladies, mais elle mérite également d'être un modèle pour la génomique des populations grâce aux vastes ressources génétiques et génomiques disponibles. L'amélioration de la tomate dépend en grande partie des introgressions d'allèles utiles provenant d'espèces apparentées sauvages.

Depuis la première diffusion d'une séquence génomique de haute qualité d'une tomate cultivée, en 2012, les génomes de plusieurs centaines d'individus cultivés et de quelques espèces sauvages apparentées ont été séquencés, permettant la découverte de millions de polymorphismes à nucléotide simple (SNPs). Leur étude a confirmé la nouvelle organisation phylogénétique et l'origine monophylétique de la section Lycopersicum du genre Solanum, composée de 13 espèces. Les approches récentes de la génomique écologique, utilisant notamment l'approche RNAseq, ont fourni de nouveaux résultats sur la spéciation et les barrières de reproduction interspécifiques. Les mécanismes moléculaires d'adaptation au stress abiotique chez les tomates cultivées et sauvages ont également été analysés et leur rôle mis en avant en tant que facteurs de spéciation et de diversification. La diversité des conditions écologiques des espèces apparentées sauvages a permis l'étude des mécanismes évolutifs et moléculaires d'adaptation au stress abiotique chez les tomates cultivées et sauvages.

Des études génomiques ont permis de clarifier les deux étapes de la domestication de la tomate et l'intensité des goulots d'étranglement dus à la domestication et à la sélection plus poussée. Les empreintes de sélection et les grandes régions génomiques introgressées des espèces apparentées sauvages ont été identifiées. Au niveau du transcriptome, il a également été montré que la domestication et la reproduction moderne modifiaient l'expression du génome, notamment pour les gènes liés au stress.

Enfin, la disponibilité des séquences génomiques et des marqueurs SNPs a permis d'étudier de grandes collections de variétés, de développer des GWAS et de faire progresser nos connaissances sur la structure du génome (décroissance du déséquilibre de liaison, distribution de la recombinaison), mais aussi de cartographier les gènes et les QTLs impliqués dans de nombreux caractères pour la sélection de nouvelles variétés.

## CHAPITRE 2 :

La domestication est un processus de sélection qui se produit sur une courte durée évolutive et qui est induit par l'homme qui laisse des traces dans les génomes des populations domestiquées. La convergence de ces changements pour des histoires de domestication indépendantes reste incertain. Il est donc nécessaire pour le déterminer, de reconstruire les changements historiques de flux génétique et de taille efficace de population, pour comprendre comment la démographie et la sélection humaine ont conjointement façonné la divergence génomique lors de la domestication. Nous avons utilisé ici un ensemble de modèles étendu basé sur des modèles de divergence démographique qui capturent la variation temporelle de la taille efficace de population et du taux de migration afin d'explorer les multiples facettes de la domestication avec le flux de gènes. Nous étudions l'histoire de la domestication de trois paires d'espèces de solanacées (aubergines, poivrons et tomates) caractérisée par des antécédents de domestication distincts, notamment l'isolement géographique du géniteur sauvage pour le poivron et la tomate et la sympatrie pour l'aubergine. Des SNPs dérivés des données RNAseq ont été utilisés pour documenter l'étendue de la différenciation génétique dans chaque paire d'espèces, et dix modèles différents ont été ajustés et comparés en fonction du spectre de fréquence allélique joint et déplié de chaque paire. Nous avons trouvé des preuves d'un goulot d'étranglement chez les trois espèces. Nos résultats suggèrent également que les quelques données historiques disponibles corroborent les périodes de domestication de ces trois espèces. Cette étude fournit donc un nouvel aperçu rétrospectif du processus d'histoire démographique qui façonne les Solanacées par le biais de la domestication et nous avons mis en avant les avantages d'effectuer la comparaison de modèles démographiques de plus en plus complexes afin de déterminer le modèle le plus adapté aux données proposées.

Nous avons cherché à déchiffrer le scénario de domestication le plus probable pour les trois paires de de populations cultivées et sauvages. Nous avons effectué une analyse comparative de plusieurs modèles démographiques de complexité croissante afin de limiter les biais induits par des hypothèses fortes. La comparaison des cultures et des populations sauvages nous a permis d'évaluer l'ampleur des changements biologiques dus à la domestication. Ces connaissances sont essentielles pour améliorer les efforts de sélection futurs et nous apportons une estimation précieuse de l'impact de la sélection humaine sur la taille de la population et le flux de gènes d'une culture efficace avec leur parent sauvage. L'inférence des scénarios démographiques de ces trois espèces est une occasion sans précédent de caractériser plus précisément la durée de chaque événement de domestication, et donc d'améliorer la déduction de l'histoire démographique qui a été supposée par des moyens indirects (histoire humaine et de culture des zones, enregistrements écrits anciens).

Résultats en bref :

L'étude comparative des inférences démographiques modélisant la domestication des trois espèces a révélé la convergence des processus de domestication dans la famille des solanacées

- Détection d'empreintes de sélection artificielles dans les génomes de Solanaceae
- Présence d’un goulot d’étranglement corroborant le stade de domestication de la culture chez les trois espèces
- Estimation du temps de divergence entre l’espèce cultivée et leur espèce sauvage apparentée :
    - Domestication de l’aubergine : 5.938-3.087 Avant notre ère.
    - Domestication du piment : 6.760-3.514 Avant notre ère.
    - Domestication de la tomate : 7.901-4.107 Avant notre ère.

Conclusion et perspectives :

- En connaissant le comportement passé de nos cultures face aux événements de domestication, nous améliorons les efforts de reproduction modernes pour soutenir la sélection future de cultures et leurs barrières innées aux conditions de contrôle humain.
- Applications possibles pour la production d'événements de domestication de novo

## CHAPITRE 3 :

La sélection consciente et inconsciente induite pendant les phases de domestication et d'amélioration variétale moderne de l'histoire de la culture a conduit à des changements phénotypiques et génétiques considérables. Les études ont porté sur les gènes à effet majeur associés à la domestication en étudiant les polymorphismes ou les niveaux d'expression génique. Dans la présente étude, nous explorons la convergence des deux processus chez trois solanacées cultivées. Pour identifier la convergence de la domestication, nous comparons la diversité génétique et les niveaux d'expression des gènes entre les accessions cultivées et sauvages dans un trio d'espèces. Nous analysons les transcriptomes de 47 génotypes, y compris des races sauvages et des races locales de tomates (Cultivées : *Solanum lycopersicum* ; sauvages *: S. peruvianum*), d'aubergines (Cultivées : *S. melongena* ; sauvages *: S. insanum*) et de poivrons (Cultivées : *Capsicum annuum* ; sauvages *: C. chilense* et *C. frutescens*). Chez les trois espèces, l'amplitude des modifications des niveaux d'expression différentielle des gènes a révélé une convergente qui est directement liée aux gènes ciblés par la sélection. En outre, la variation de l'expression des gènes était significativement corrélée à la variation de la diversité des nucléotides chez les trois espèces. Alors que nos analyses transcriptomiques ont confirmé les changements d'expression de nombreux gènes liés à la domestication, la nouveauté de notre étude réside dans la convergence des empreintes de domestication agissant à la fois sur la diversité des nucléotides et sur l'expression des gènes.

Dans le troisième chapitre, l'hypothèse est une modification convergente de la diversité et de l'expression des gènes lors de la domestication. La comparaison des accessions relatives des cultures et des espèces sauvages relatives a permis d'estimer les différences d'expression génique et de détecter les empreintes de sélection génomique. Les annotations des gènes ciblés (sélectionnés et exprimés de manière

différentielle) ont permis d'identifier les processus biologiques modifiés lors de la domestication. L'hypothèse repose sur les orthologues partagés au sein du trio d'espèces et leur modification. Nous émettons l'hypothèse que les mécanismes de régulation et d'adaptation déclenchés par la domestication des espèces cultivées sont convergents. Par conséquent, pour les trois processus de domestication indépendants, il est prévu de mettre en évidence des changements parallèles induits dans les cultures par rapport à leurs parents sauvages.

Résultats en bref :

L'étude des orthologues a mis en évidence une convergence des modifications moléculaires chez les trois espèces,

- Au niveau génétique :
    - Relaxation de la sélection : initiation de la transcription, initiation de la traduction et tolérance aux stress abiotiques
    - Sélection de la direction : croissance de la plante et développement du fruit
- Au niveau de l'expression des gènes :
    - Régulation négative (baisse) : régulation des réponses abiotiques et de la tolérance à la sécheresse
    - Régulation positive (hausse) : croissance des plantes, expansion des cellules, croissance des feuilles, développement des fruits et maturation

Conclusion et perspectives :

- Les voies touchées par la domestication sont mondiales et ont donc un impact sur plus de voies polygéniques que de gènes locaux.

- L’approfondissement de l’étude pourrait permettre de détecter des gènes spécifiques co-exprimés impliqués dans la domestication
- Applications possibles pour retrouver des caractéristiques adaptatives telles que la tolérance à la sécheresse chez les populations sauvages
- Véritable souci de conserver les populations sauvages en tant que sources de diversité

- Quelles étaient les espèces sauvages parentes des espèces cultivées actuelles ?

La première question concernant le géniteur sauvage de chaque espèce semblait avoir une réponse lorsque les échantillons ont été choisis. Grâce à nos analyses, nous pouvons vérifier une fois de plus que *Solanum melongena* a été domestiqué à partir de l'espèce sauvage *S. insanum* et que *S. pimpinellifolium* est le géniteur sauvage de *S. lycopersicum*. *Capsicum annuum* var. *glabriusculum*, supposé géniteur sauvage du piment cultivé, doit être plus étudié car nos résultats démontrent la forte disparité de l'espèce. Le poivron cultivé a sûrement été domestiqué à partir d'une sous-espèce de *C. annuum* var. *glabriusculum* mais la structure de l'espèce doit être éclairci pour permettre des analyses plus poussées de sa domestication.

- Quel impact a eu la domestication sur les génomes et les transcriptomes des espèces cultivées ?

L'étude s'est concentrée sur des données RNAseq, donc sur la partie exprimée du génome. En comparant les espèces cultivées à leur sauvages proches, chez les trois espèces de Solanacées, nous avons pu détecter des changements considérables dans la diversité nucléotidique et dans le niveau d'expression des gènes, changements dus à la domestication. La corrélation entre ces changements (diversité génétique et variation de l'expression des gènes) a révélé la convergence des mécanismes de régulation à l'échelle du génome et du transcriptome lors de l'adaptation à la domestication. Des études complémentaires sur des études métabolomiques pourraient conduire à une compréhension plus complète de chaque niveau de régulation moléculaire.

- Quels étaient les gènes et les voies ciblés par la sélection ?

L'étude des orthologues a révélé des gènes communs ciblés par la domestication chez les trois espèces ce qui a confirmé une convergence de sélection due à la domestication. La domestication a eu un impact positif sur les traits liés au syndrome de domestication tout en modifiant négativement les voies impliquées dans la tolérance au stress et la résistance aux maladies.

- Et finalement, comment peut-on utiliser les espèces sauvages apparentées à nos espèces cultivées pour recouvrer des traits perdus et améliorer les cultivars modernes ?

Tout en identifiant les gènes ciblés par la domestication au sein de l'espèce cultivée, avec les analyses comparatives, l'espèce sauvage apparentée révèle son potentiel en tant que ressource génétique pour la récupération des traits de caractères perdus lors de la sélection associée à la domestication. La diversité génétique qui reste chez les espèces sauvages apparentées est une occasion d'améliorer considérablement les faiblesses des espèces cultivées ou des cultivars modernes, par exemple, pour récupérer les résistances aux maladies ou la tolérance au stress environnemental.

Ce travail confirme la nécessité de conserver les espèces sauvages apparentées et les races locales dans des collections de base plus représentatives. En particulier, dans un contexte où les races locales, qui ont été maintenues pendant des milliers d'années, disparaissent lentement avec l'exode rurale des populations autochtones, telles que les populations autochtones amérindiennes qui étaient le centre de conservation de la plupart des anciennes races locales (Smith et al. 1992).

Cette thèse porte sur les caractéristiques communes et divergentes des espèces cultivées et de leurs espèces sauvages apparentées. Les méthodes comparatives sur les données RNAseq ont permis de détecter efficacement les changements

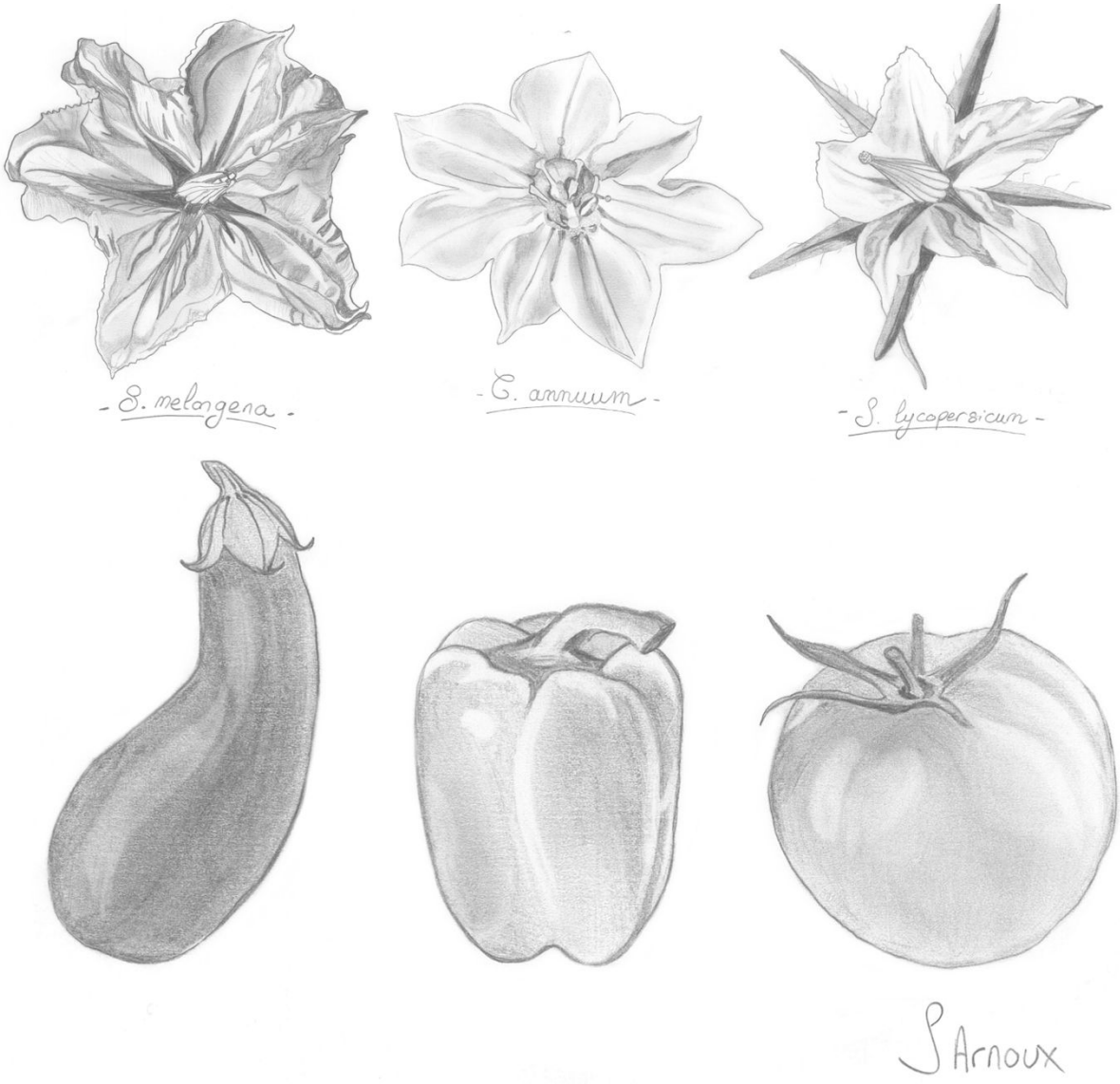
d'expression et de polymorphisme nucléotidiques dus à la domestication. Les résultats présentés ici fournissent, à partir des inférences démographiques, une estimation de la durée de la domestication et, à partir des analyses transcriptomiques, un aperçu des conséquences génétiques et transcriptomiques du processus de domestication. Les deux articles confirment la perte de diversité génétique au sein des espèces cultivées.

Globalement, ces résultats mettent en exergue les opportunités d'améliorations futures liées à la perte de gènes d'adaptabilité au sein des espèces cultivées qui sont encore présents dans les espèces apparentées sauvages. Depuis 1945, plus de 30% de l'augmentation du rendement des cultures peut être attribuée à l'utilisation d'espèces apparentées sauvages dans les programmes de sélection végétale (Pimentel et al. 1997). Dans ce contexte, ces travaux confirment la nécessité de soutenir la conservation des espèces sauvages apparentées dans leurs environnements sauvages et dans les centres de ressources génétiques. Les analyses des changements dus à la domestication révèlent notamment des traits d'intérêt conservés dans les pools de gènes sauvages. Les races locales et les espèces sauvages pourraient être utilisées pour compléter la population de référence pour de futures analyses à l'échelle du génome afin de détecter les régions susceptibles d'être améliorées. Les régions détectées pourraient ensuite être introgressées dans les cultivars modernes pour améliorer leur tolérance aux stress et leurs résistances aux agents pathogènes. En parallèle, les variations épigénétiques et métabolomiques sont toutes deux sources de diversité phénotypique. Ainsi, en utilisant la biotechnologie émergente, la modification de la régulation des gènes et de la composition métabolique pourrait générer des améliorations essentielles du rendement et des caractéristiques nutritionnelles des espèces cultivées (Harrigan et al. 2007; Gallusci et al. 2017). À terme, les efforts de sélection moderne augmenteraient considérablement les données phénotypiques et génotypiques permettant l'utilisation de la sélection génomique. Cette méthode connecte les phénotypes et les génotypes connus et les utilise comme paramètres a

priori pour modéliser et prédire les phénotypes à partir des génotypes (Morrell et al. 2011).

Ces travaux fournissent une plate-forme de base aux programmes de sélection modernes en fournissant une liste de gènes orthologues qui ont été ciblés lors de la domestication chez les trois espèces de Solanacées. La convergence de ces changements offre une opportunité considérable d'utiliser les connaissances transversales pour améliorer les cultures, par exemple en utilisant la manipulation des gènes trans-espèces (Bastet et al. 2018). C'est particulièrement le cas dans la famille des solanacées qui possède une grande synténie offrant la possibilité de transférer des connaissances à d'une espèce à l'autre (Rinaldi et al. 2016).

Ce travail de thèse confirme ce que Darwin suggérait déjà il y a plus de cent ans: étudier le processus de domestication a un grand potentiel tant pour mieux comprendre la sélection artificielle et l'évolution convergente que pour apporter des informations précieuses pour l'effort de sélection et l'amélioration moderne.


- S. melongena -
- C. annuum -
- S. lycopersicum -
S Arnoux